APL/JHU CF-2885

August 1960

Copy No. 58

PAPERS ON OPTIMIZATION THEORY

Presented at
the APL
Associate Staff Training Program
in 1959

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
Silver Spring, Maryland

OPERATING UNDER CONTRACT NOrd 7386 WITH THE BUREAU OF NAVAL WEAPONS, DEPARTMENT OF THE NAVY

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

PAPERS ON OPTIMIZATION THEORY

Presented at
the APL
Associate Staff Training Program
in 1959

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

PREFACE

This series of lectures on optimization techniques was originally prepared for presentation to the participants of the 1959 Associate Staff Training Program. Its objectives were two-fold: (a) to introduce engineers to the usefulness of mathematical techniques and mathematicians to the applicability of their techniques to concrete engineering problems and (b) to introduce all the trainees to the very important concept of optimization--a concept that currently pervades virtually every area of advanced missile engineering.

Because these lectures do form a sound, basic introduction to the important field of optimization techniques, they are being issued as a CF report to provide wider distribution of this material as a reference work for Laboratory personnel.

V. M. Root
Training Program Supervisor
August 1960

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

TABLE OF CONTENTS

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

ON THE CONCEPT OF OPTIMIZATION

"Fulsomely dedicated to Winnie-the-Pooh,
who didn't know a minimum from a Heffalump's trap."

by

A. G. Rawling

I. INTRODUCTION

A somewhat negative approach to the principle of optimization was first scribbled centuries ago, among other phrases, on the walls of an ancient Roman bath house. Referring to a choice between two Roman aspirants for emperor, it read -- "De duobus malis, minus est semper aligendum" [Of two evils, always choose the lesser].

Since that time optimization has been carried on under many guises. In physics, many different minimal principles have been enunciated, describing natural phenomena in the fields of optics and classical mechanics. The field of statistics contains various principles termed "maximum likelihood," "minimum loss," and "least squares," while economics contributes "maximum profit, minimum cost, maximum use of resources, minimum effort," in a coherent effort to increase the long run capital gain in some manner.

Enlarging our viewpoint to include the most general aspects, we note that many operational problems are of this sort.

1. They have a variety of acceptable solutions (by some specific criteria of acceptability).

2. Among these solutions one wishes to select the best or optimal solution (by some specific criteria of being best or optimal).

Thus, one formulates the problem mathematically with the twin objectives of providing an accurate description and also manipulating the mathematical model to obtain an extremum.

An extremum, or extreme value, is a value of a function which is either a maximum or a minimum. Optimum is that particular type of extremum desired for the problem.

II. FULL, INCOMPLETE, AND SUBOPTIMIZATION

Full optimization of a given problem can be quite an extensive undertaking. It requires

1. simultaneous consideration of all possible alternatives at all levels;
2. consideration of the probable impacts of all events not under the optimizer's control; and
3. maximization, subject to possible constraints, of some utility function or measure of effectiveness.

Failure to achieve full optimization results in "incomplete optimization," of which a special variety is termed "suboptimization."

Suboptimization is a case of optimization for one phase of an operation or a problem, without including every factor which has an effect, either obvious or indirect. Frequently, it consists of merely reducing the number of objectives.

The suboptimization approach is useful when neither the problem formulation nor the available techniques permit one to obtain a reasonable answer. In most practical cases, suboptimization is the only resort in solving the problem. Although a full optimization is not obtained, it at least provides a rational technique for approaching the optimum.

Suboptimization is often necessary because of economic and practical considerations, the finiteness of time, and the difficulty of obtaining sensible answers in a hurry.

However, there is a major fallacy to be guarded against. Suboptimization of all elements does not necessarily ensure attainment of full optimization, i.e., an over-all optimum. For example, in a large business, the sales force endeavors to increase sales of all items, although the profit on each item may differ.

The production group resists changes to new products. The comptroller wishes to reduce inventory level so as to

free capital and decrease storage costs, etc. It is not difficult to suboptimize all these divisions separately, so that each is running smoothly and effectively, but a more painstaking effort is necessary to balance the tendencies of different parts of a large organization and ensure they all mesh together.

System design* is sometimes described as the process of attaining a full optimization. The need for system design arises from this very fact that suboptimization of all components does not necessarily improve system performance.

An excellent example of this has actually appeared in the following concrete situation involving target tracking by a homing missile.

The use of doppler information to aid range gate tracking has long been known. In obtaining the optimum over-all system, usually the range and speed tracking loops are separately optimized according to some criterion and then the speed gate is connected to the range gate to further reduce the range error.

The resulting system may approach the optimum over-all system; however, the best use of the additional information consisting of the correlation between speed and range has not been made.

The optimum circuit consists of two optimum loops for range and doppler information tied together in an optimum fashion. Consequently, we require optimum gains connecting one circuit to the other as well as gains in the individual circuit loops.

III. OPTIMIZATION TECHNIQUES AND CONSTRAINTS

We are all familiar with the optimization problem of finding, among all the paths between two points in the plane, the shortest path. It can be shown that it is a straight line.

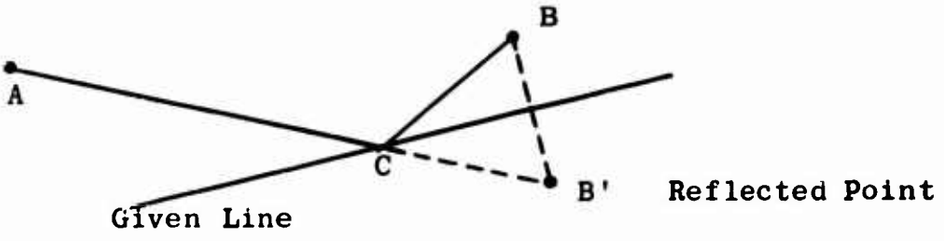
Consider the extension, where the two points lie on the same side of a given straight line, and the problem is to

---

*"A system is an integrated assembly of interacting elements, designed to carry out cooperatively a predetermined function." - Dr. R. E. Gibson.

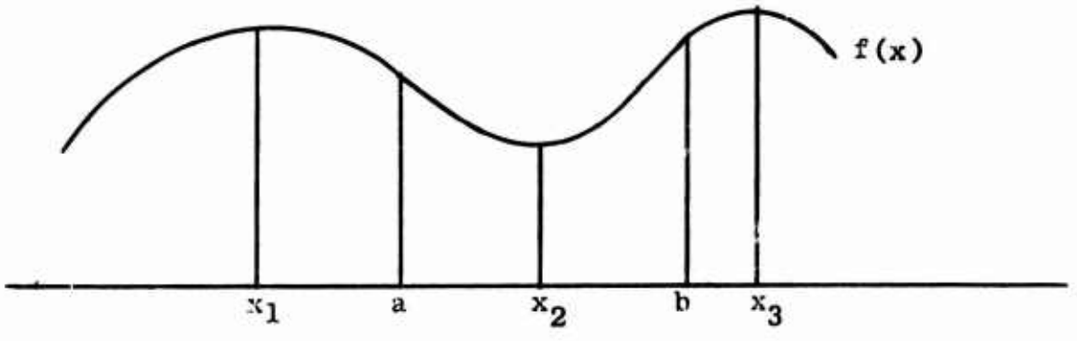
determine the shortest path between the two points that touches the given line. This added condition is a constraint on the problem.

(The solution is not difficult in this case. Reflect one of the two points across the line, and connect the other original point and the reflected point with a straight line, as shown. Then ACB is the shortest path satisfying the constraint.)



Optimization of some function of a number of variables subject to boundary conditions which limit the variables' range is of greater importance and frequency of occurrence than optimization without constraints. We usually are not interested in relative maxima or minima, but rather maxima or minima over a prescribed range. This entails additional restrictions.

For example, in the range $a \leq x \leq b$,



there is a maximum at x = b, but not at x_1 or x_3 which lie outside the range of interest. (This so called end point maximum will be discussed more fully later.) There is a minimum at x_2 within the range.

The difficulties in the more interesting problem of optimization subject to constraints occur in many cases where the maximum or minimum cannot be obtained by ordinary differentiation because either it does not lie within the region defined by the constraint set or else the derivative is discontinuous inside the region.

If the function to be optimized is continuous, then the extremum lies either in the region or on the boundary. If it lies within the range, regular methods of finding maxima and minima apply as if the inequalities were not present.

Mathematical constraints are of two types: equations and inequalities. In a typical problem, the constraints may be all one or the other, or a mixture. Considerable difference exists in the techniques applied to problems in which all constraints are equations or else all inequalities.

If all constraints are equations, then in principle each constraint equation can be solved for one of the variables and this substituted to reduce the dimension of the problem by one. For example, to find the extrema of f(x, y) subject to the constraint $\phi(x,y) = 0$, we solve for y from $\phi(x, y) = 0$ to obtain $y = \psi(x)$. Then substitute into $Z = f(x,y)$ and extremalize

$$Z = f\left[x, \psi(x)\right]$$

as a one-dimensional problem.

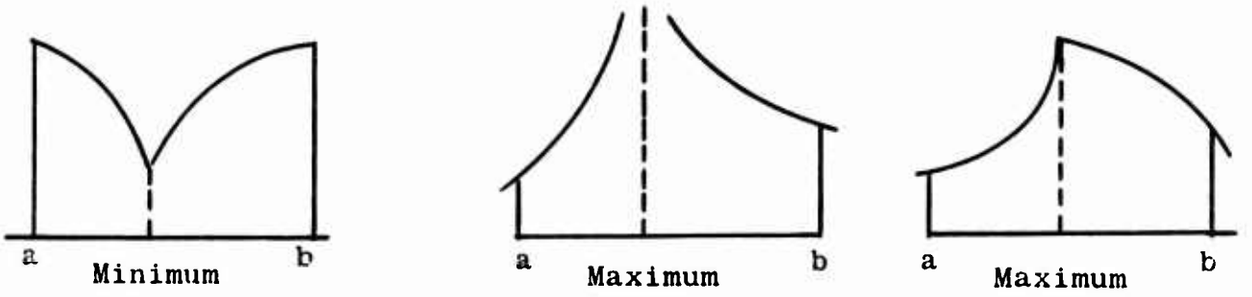
An inequality constraint does not make it possible to "eliminate" any variables; it merely restricts the range of variability in one dimension. Thus, if the extreme should be on the boundary or end point of the range, a different method for extremalization is necessary.

An end-point technique is as follows: Given a function f(x) defined for $a \leq x \leq b$, and let f(x) have a derivative.

Then f(x) has a maximum at x = a if $f'(a) \leq 0$, and a maximum at x = b if $f'(b) \geq 0$.

It will have a minimum at x = a if $f'(a) \geq 0$, and a minimum at x = b if $f'(b) \leq 0$.

It will occasionally happen that f'(x) becomes discontinuous for some isolated value of x, and if the discontinuity is accompanied by a change of sign as x increases thru the value in question, we shall have a maximum or minimum, as shown.



These remarks pertaining to one variable are generalizable to two or more variables, where the boundaries become curves and surfaces and their intersections. The difficulties of testing increase also.

Actual constraints, which will be discussed elsewhere under several examples, are many and varied. They can include: negative production not allowable, maximum limits on storage capacity, and production capacity in economic problems. Nonlinearities in physical systems, such as saturation and limiting, can be described mathematically in the form of inequalities. Noise or uncertainty is often describable in terms of a probability distribution or power spectra. Competitive or game theory aspects, such as opponents' strategies, must be considered in many cases.

Physical constraints, involving weight and size are by no means minor. For example, an optimum cargo ship hull design problem separates into two classes:

1. Designs in which size (displacement) is fixed, as determined by available power.

2. Designs in which dimensions (particularly length) are dixed, as determined by practical considerations of port facilities, such as dock size.

Different optimum solutions are to be expected in each case.

IV. THE CRITERIA OF OPTIMALITY

A major difficulty in any optimizing problem is that of selecting a criterion. What is the criterion in terms of which the outcome is to be judged?

In game theory, the sensible object of a player is to gain as much from the game as he can, safely, in the face of a skillful opponent who is pursuing an antithetical goal. The classical criterion for optimizing the design of a mechanical device has been to maximize the output energy for a given input energy. In the case of vibration of a continuous system, the optimum damping value gives the least resonant amplitude.

In choosing an optimization criterion, several aspects must be considered.

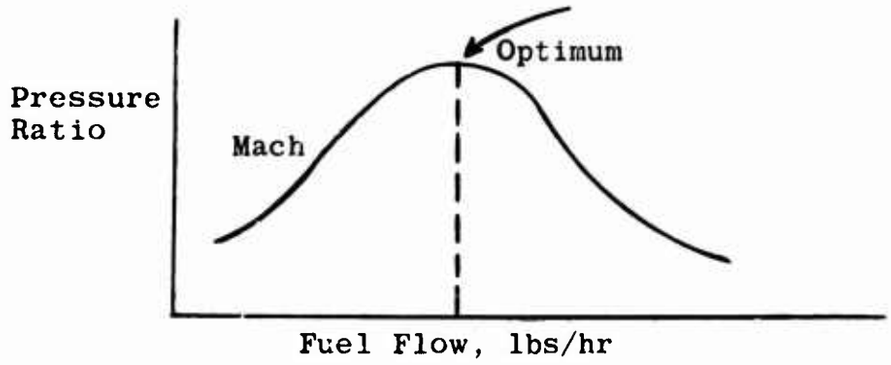
First, viewpoint is important. For example, consider a nonproduction line type of manufacture, i.e., the product is generated in discrete batches. From the viewpoint of reducing storage costs, we might ask "How can inventory level be reduced?" But a wider viewpoint is represented by the question "What is the optimum size of inventory with respect to making a profit?" From the savings resulting from larger batch size with less frequent production, it might be that the inventory size should be increased.

Constraints are also important in criteria selection. For an airborne digital computer, weight and size are fixed, so that any criterion for optimal choice among several computers must include the question "Does it fit?"

There may well be a multiplicity of criteria to plague the optimizer. They must be reduced in some way to manageable proportions by grouping them, inter-relating them, or just discarding them. Conversely, one of the main reasons that criteria selection is difficult is the fact that one usually deals with either incomplete or suboptimization. In most cases, however, the performance criterion is the first to be examined.

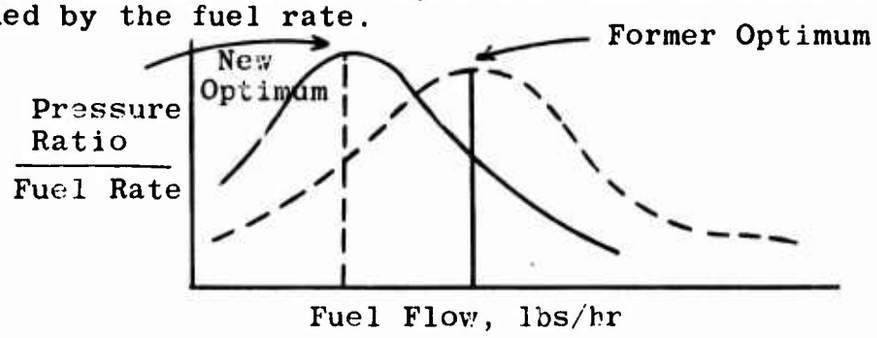
Consider the application of optimalizing principles to the field of automatic controls for an aircraft jet propulsion system.

One such optimalizing control meters the input fuel flow to an engine in such a manner as to produce the maximum compressor output pressure over a range of flight conditions.



This indicates the goal is to maintain delivery of thrust without consideration of fuel consumption.

Clearly, if the economical use of fuel is the primary purpose, then the output might have been the ratio of pressure divided by the fuel rate.



Here the optimum point (a maximum) has shifted to a lesser fuel flow. Which optimum is better? No statement is possible. Each is an optimum for its chosen criterion.

To sum up the essence of this section, the optimal ox consists of

all meat to the gourmand,

all hide to the shoe seller,

and all hoof to the glue maker.

Therefore, always define in advance your chosen criterion of optimality and the constraints.

V. VARIOUS EXAMPLES

Optimum Coding of Information --- Communication and information theory abound with the word "optimum." Having determined a cost in energy, time, or money of transmitting each one of a set of symbols (e.g., Morse code), what is the optimal code using these symbols which will transmit a given amount of information at the least cost, or will transmit information at a given rate for the least cost per unit time? (If noise is present, the cost of each symbol is effectively increased.)

Thus, in an optimum coding process, we try to produce a set of symbols (each which may take different times to transmit) to be sent over the communication channel so that they will all occur independently and with equal frequency. This will permit a message to be encoded in such a way as to utilize the fixed channel capacity in an efficient manner.

Sometimes optimum code means a minimum-redundancy code. This is a code which, for a message ensemble consisting of a finite number of members and for a given number of coding symbols, yields the lowest possible average message length.

In general, the optimum code alone may not be identical with the optimum code when channel characteristics enter as constraints.

Optimum Programming of Computers - -The speed of a storage device, such as a computer memory, is measured by its access time, the time required for either reading or writing access to the first location required. Access can be random (each bit available within the fixed access time) in the case of magnetic cores, or it can be cyclic (in which the access time depends on where the bit is in the cycle) as in the case of a magnetic drum rotating past a group of magnetic reading/writing heads.

If a computer program is stored wordwise in a sequential manner around the rotating drum, the time delays involved in reading or writing in one memory location prevent the next location from being processed immediately, and the drum must

spin around again before the next word location can be utilized, and the time to run the computer program is considerably extended by this waste of time.

Optimal programming consists of interweaving the locations, i.e., arranging the program to permit space between consecutive storage locations. This will permit consecutive reading despite the control time lags. In this way the time of running is reduced. Computers with cyclic memories (drum, discs or delay lines) require such optimal (or minimal latency, as it is also known) programming so as to minimize the computer time required for a given program. (A disadvantage is the widescattering of orders through the drum memory. This may cause considerable difficulty when several programs have to be fitted together.)

A first approach consists of spacing the words a fixed distance apart, regardless of the unequal length of time different commands require for execution. A distinct improvement results when the orders are spaced a variable distance apart, so that a location on the drum tends to be passing the magnetic reading head when it is needed for access.

Optimum programming usually does pay, but not always. In order to pay, the computer time saved must exceed the additional programming time required for optimizing (subject to the qualifying discussion at the end of this paragraph). It will always pay if (1) the program is to be used repeatedly in processing large volumes of data and if (2) the programming can be done before the numerical data for the problem is available. It may pay if the problem is a long, nonrepeatable one, with a large amount of data.

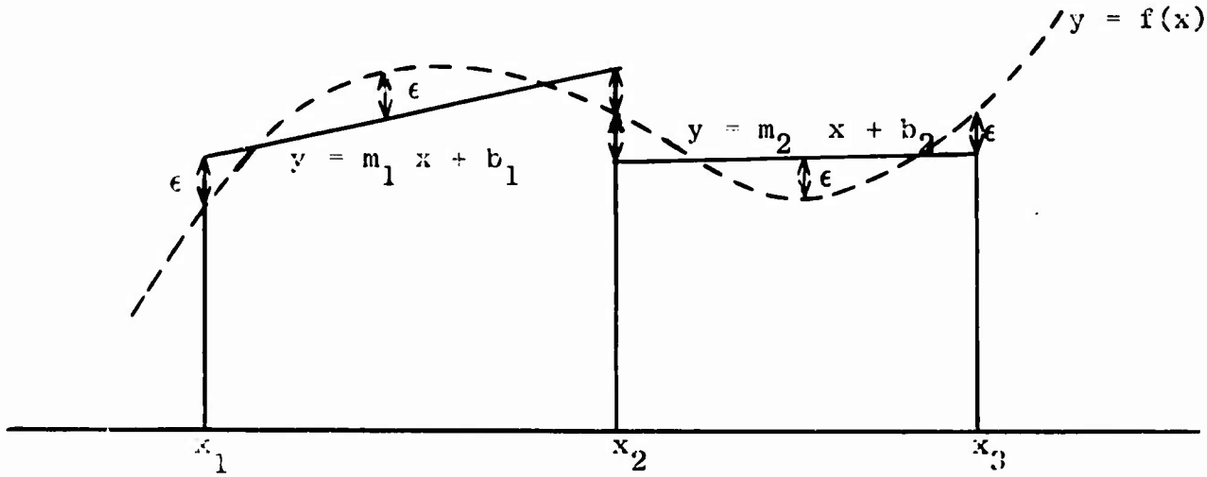
(Qualifying discussion: From a cost accounting viewpoint, it is natural to minimize the large expense of using digital computers by emphasizing the importance of "minimum machine time" concept as a programming philosophy. Since the cost of one hour machine time is approximately equal to one week of programmer's salary, it might seem reasonable for a programmer to spend one week's effort "optimizing" the program to save more than one hour running time. However, more often than not, several scientists or engineers are held up for a week awaiting an answer before they can take action on it. In the broad view, it may be less expensive to get answers to the sponsor before he forgets the problem, even if it costs more in machine running time.)

Optimizing the computer program can be both tedious and

valuable. Can a machine be programmed to optimize its own programs? Yes -- but certain compromises are necessary, principally the problem program and the optimizing program must both fit in the memory at the same time. This limits the size of the problem that can be optimized.

Optimum Interval Tables --- Instead of storing tables of function values in a digital computer, or printing subtables of differences on each page, polynomial approximations of the tabled function can be made, valid throughout the tabular region, with a given allowable error. These polynomial representations can then be evaluated to reconstruct the function for any argument in the region, and thus include interpolation as well as tabulation.

Thus, we can assign in advance both the degree n of the approximating polynomial and the maximum allowable error ϵ. For example, in the sketch shown below, the curve $y = f(x)$ would be replaced by straight lines ($n = 1$) such that the error everywhere is less than a prescribed ϵ (this fact determines the length of each subinterval).



Now for an interpolation polynomial of degree n and allowed error ϵ, the computer merely stores the coefficients m_i, b_i of each interval and evaluates the linear polynomial corresponding to the argument x lying within the interval.

This process, termed optimum interval interpolation minimizes the number of subintervals needed over the entire range. Function tables so constructed are called optimum interval tables.

VI. CALCULUS OF VARIATIONS

How does an optimization problem in the calculus of variations differ from an extremal problem of ordinary calculus?

In the latter case, we are given the function $y = y(x)$. A simple problem in ordinary calculus is to find a value x which yields a minimum or maximum value of $y = y(x)$.

One step removed from finding the extremum of a differential function is the basic problem of the calculus of variations: to find a function $y = y(x)$, instead of a value of the variable x, which makes a certain definite integral

$$I = \int_a^b f(x, y, y')\, dx$$

take on a maximum or a minimum. Note we cannot integrate directly, because y is not known as a function of x, hence $f(x, y, y')$ is not known as a function of x. Thus, the ordinary methods of solving maxima and minima problems do not apply.

Geometrically, the calculus of variations deals with the problem of finding paths of integration for which integrals admit maximum or minimum values. Solutions may be either continuous or discontinuous in the first derivative, or both, i.e., "corners" may exist in the path, occurring at the junction of different continuous arcs.

Extensions to the basic problem include the presence of higher order derivatives in the integrand, multiple integrals involving partial derivatives in the integrand, variable limits of integration, and constraints represented by the requirement that another integral, involving the same variables, has a constant value. Classical applications of the calculus of variations include the problems of finding the minimum surface of revolution, the maximum solid of revolution, least action, solids of minimum resistance with and without the constraint of given volume. (A recent technique, called "Dynamic Programming" is a powerful computational approach to both

classical and nonclassical problems in the calculus of variations.)

VII. OPTIMIZATION AND ROCKETS

The field of rocketry provides an exceedingly ripe area for both optimization criteria and mathematical techniques involving the calculus of variations.

For example, consider the problem of specifying the rocket trajectory. Different methods of control exist in flight. The thrust may be varied in magnitude, and often the thrust direction is variable. These changes affect the flight trajectory, there may exist some single goal, such as attainment of long range with minimum expenditure of propellent and structural weight, or attainment of some altitude in a reasonable time. It is of interest to find out how to adjust the available control so as to optimize the trajectory in the sense of maximizing or minimizing some function such as range or time subject to constraints such as fuel consumed or altitude achieved. Since the controls can usually be varied at will over a continuous range of values, such trajectory problems belong to the calculus of variations.

Over the past years, there have been numerous applications of the calculus of variations (as well as other techniques) to optimization of rocket problems. Some of the many topics published include the following:

1. Either maximize the range, altitude or some other property for a given fuel consumption. or specify such a property and seek to minimize the fuel consumed.

2. Program the exhaust velocity in an optimal manner so as to provide the most efficient utilization of the fuel.

3. Determine the optimum thrust direction of a rocket fired from a fighter plane pursuing a constant velocity target in order to maximize the initial missile-target range.

4. Determine what value of payload that will give maximum kinetic energy for a rocket of fixed structural and propellent weights. An optimum ratio of payload to structure exists for every value of the propellent ratio.

5. Obtain optimum staging techniques for a multi-stage rocket with different construction parameters and propellent specific impulses in each stage. We wish to optimize the configuration insofar as performance is affected by changes in the number of stages, and redistribution of the fuel and structure weight among the stages. An optimization criterion might be maximum burnout velocity for given take-off weight.

6. Seek an optimum nozzle area ratio for rockets operating in a vacuum which will provide maximum performance for the stage in question. (Here the specific impulse of the propellent increases monotonically with increasing ratio of exit area to throat area, but the increased weight of larger nozzles degrades the performance.)

VIII. OPTIMUM EFFORT OR SEARCH

Suppose that an object is somewhere in a given area, and that its probabilities of being in the various possible positions are known.

Suppose, further, that a limited total amount of searching effort (or time) is available.

Finally, assuming that the law of detection is known, the chance of finding the object when a given amount of search is carried out in its vicinity is determined.

The major problem is to find the optimum manner of distributing the available searching effort: the one which maximizes the chance of finding the object (i.e., remember detection is not certain. Rather, detection is an event which may have any probability between zero and one associated with it).

If the object is equally likely to be anywhere within a certain area, the problem is straightforward. The search effort is evenly laid out over as much of the area as we can search.

But if the chance that the "enemy" is present varies from area to area, the problem can become quite difficult. As an example, if the enemy is twice as likely to be in one area than another, and only a small amount of search area is available,

all this effort should be spent in searching the more likely area.

Additional by-products are such questions as "What would be a good course of action? Or several courses, possibly?" "What is the best way for the opponent to hide?"

A more specialized problem is as follows. A target is known to be in a large volume. Using a pulse radar, the volume is searched by scanning systematically over the entire space. Assume further that the search situation can be described by

1. the size of the volume the target is known to be located in;
2. the strength of the return signal.

The parameters to be optimized are:

1. the return threshold;
2. the size of the unit search interval;
3. The time spent examining each interval.

A related problem is the optimum acquisition procedure, i.e., to minimize the average time to acquire a target with a radar while constraining the cumulative probability of a false alarm.

IX. SETTING THE OPTIMIZATION PROBLEM

If we now consider systems primarily, henceforth, we can list five conditions which largely determine any problem in system optimization:

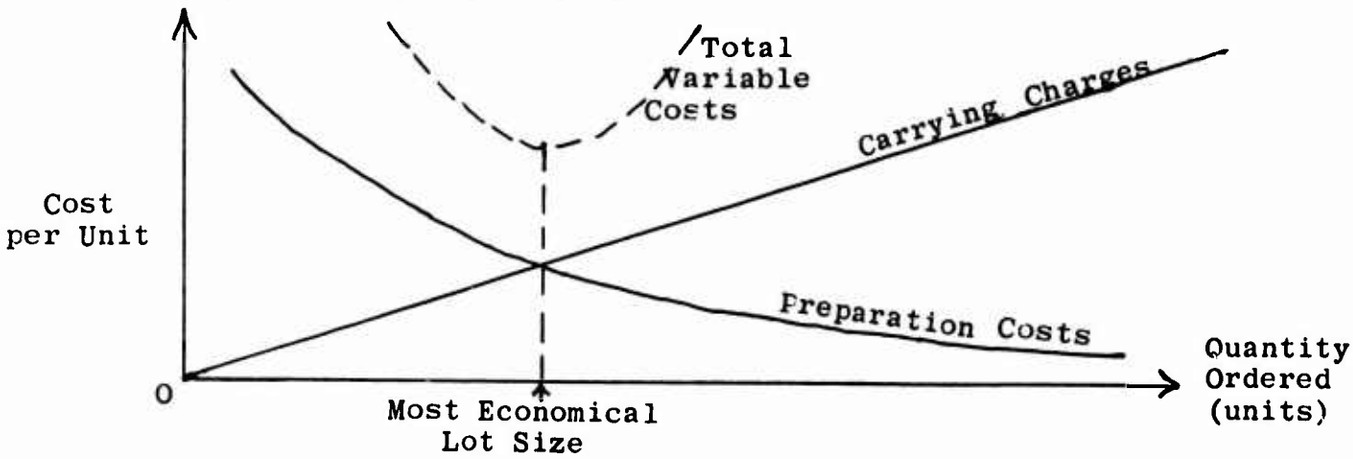
1. Purpose of the system;
2. Nature of the inputs;
3. Criterion of goodness of performance to be used;

4. Freedom of choice to be allowed in design;

5. In practical problems, cost of the system in a generalized sense must be included.

Whenever these five conditions are specified, some kind of optimization problem is defined, although it may be such that the problem has no solution at all, or no best solution, or no unique best solution.

The question as to whether an optimum exists or not is dependent in general on the existence of at least two opposing functions in the system. For example, in the determination of economic lot size, the total variable costs are the sum of carrying charges plus preparation costs, as shown below:

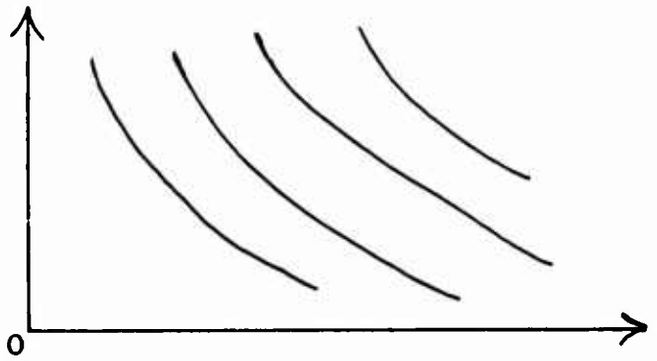


The most economic lot size occurs where the total variable cost is a minimum.

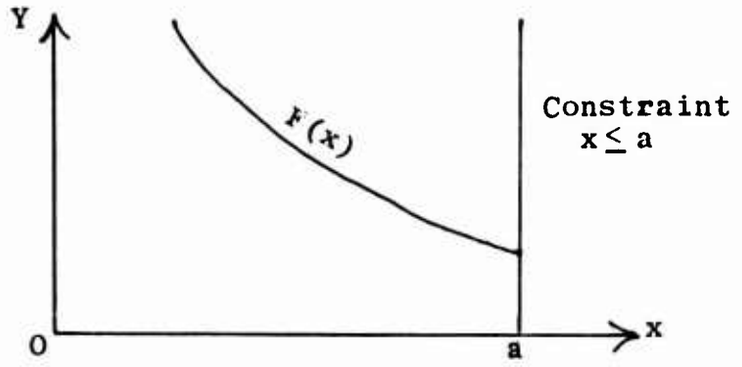
Another example: The velocity required to leave a satellite orbit about one planet and proceed to another planet depends upon the radius of the satellite orbit. The larger the satellite orbit, the weaker is the gravitational attraction of the home planet. This has a tendency to reduce the velocity required to go to another planet.

But the larger the satellite orbit, the lower is the circular velocity of an object in that orbit. This has a tendency to increase the additional velocity required. Therefore, there is an optimum satellite orbit of departure of arrival to go from each planet to each other planet.

On the other hand, if no opposing functions exist, there may be no optimum.

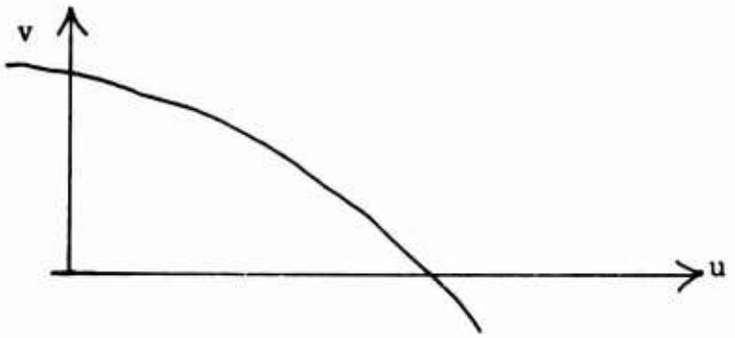


However, the presence of bounds on the variables as constraints may serve to introduce the so-called end-point or boundary optimum, as mentioned earlier.



In the above case, we see a minimum occurs on the boundary at $x = a$.

Other nonmathematical difficulties may beset us. Suppose we have two given functions $u(x, y, z \ldots)$ and $v(x, y, z \ldots)$ to make as large as possible (e.g., quality and profit), but we are thwarted by the fact that increasing one of them decreases the other.



Two traditional methods are available:

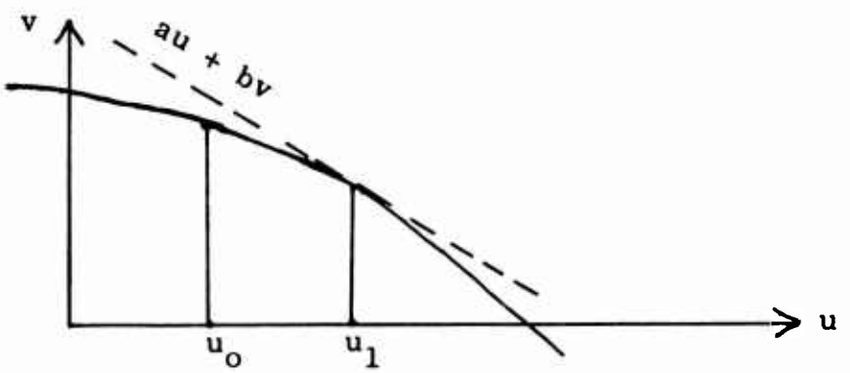
1. Fix one variable, say u, at that particular value u_o the least that will be tolerated, and maximize v subject to this constraint.

2. Take a weighted average of the two variables

$$au + bv$$

where a and b are non-negative and sum to unity, and minimize this mean, to give an optimum u_1, v_1.

In general, the two optima at u_o and u_1 will not coincide.



The point is, the choice of either u_o, (the least tolerable u) or the weights a, b (the balance between u and v) is <u>not</u> a mathematical but an executive decision.

X. OPTIMIZATION OF LINEAR CLOSED LOOP SYSTEMS

A fundamental function of feedback controlled systems is to maintain certain variables (called errors) at constant or minimum values, under a particular set of conditions which include:

1. dynamical nature of the process to be controlled;
2. difficulties in accurate measurements;
3. servo limitations as to power and saturation;
4. random disturbances and inputs;
5. noise.

Any effort to establish a routine design procedure of such a system requires specific criteria defining the optimum system.

We are fortunate. There is no dearth of criteria. Among the multitude are such ones as zero error, response specified by a model, variable damping dependent on error size, minimum lead and bandwidth, transient response characteristics, optimum impulse response, minimization of the effects of disturbances, maximization of system output, minimization of the square of weighted noise and dynamic error, maximization of load torque at constant speed, stable equilibrium, any of a host of integral-of-the-error type of criteria, including minimum mean square error.

This great number of various criteria can be reduced in general to three basic sets of performance criteria, existing at present in problems of servo mechanism design. They are:

1. Stability;
2. Steady-state conditions;
3. Transient response characteristics.

In general, a device which is not stable will not be used. (In some multiple loop feedback systems, sometimes an inner loop by itself might be deliberately designed to be unstable in order to achieve some desired benefit, but when the subloop is embedded properly, the over-all multiple loop system must be stable.)

Transient response refers to the manner in which the device arrives at a steady state. Investigation of the system for the case of transient (or nonsteady-state) phenomena usually, but not always, requires approaching the problem in the time domain.

Optimum transient response may be defined as that response to a step input which regains position correspondence in minimum time and with prescribed limitations on overshoot [e.g., for a nonlinear relay servo we might prescribe no overshoot at all]. Other definitions also exist.

The phrase "optimum adjustment" signifies the problem of setting the adjustable parameters of a control loop so that the control action resulting after a disturbance will take place in the best possible manner.

Of course, no such optimum adjustment will be universally applicable, because it is always based upon the criterion used to define optimum control action. The choice of such a criterion is rather subjective and depends upon each application.

Steady-state conditions refer to the error (between input and response) which remains after all transients die out.

Determination of the optimum transfer function in the steady state requires working in the frequency domain.

There are two distinct methods of design which we will encounter now in automatic feedback control and later in optimal filtering. They are:

1. Parameter-optimized, or fixed configuration, or relative optimum.

2. Variation-optimized, or free-configuration, or absolute optimum.

Parameter-Optimized---This is a simplified method in which it is necessary at the outset to decide the interconnection of the elements or blocks to be employed. The procedure to optimize parameters in the proposed (or existing) system consists in varying the parameter values that are not specified previous until the system gives minimum mean-square error, i.e., after first obtaining an expression for the mean-squared error in terms of the parameters of the system, the optimum parameter values may be determined, using ordinary minimizing techniques of calculus.

The success depends to some extent on the wisdom of the initial choice of interconnections and elements. There must always remain some doubt as to whether some other interconnection might not lead to a better result.

Variation-Optimized-- This method relieves the designer completely of the onus of an initial choice of elements. It is based upon the calculus of variations and when the other fixed elements in the system have been specified, it allows no choice in the remaining elements, but gives directly an absolute optimum which cannot by any linear means be improved upon.

Such systems may be difficult to realize in a practical form since the method takes no account of practical convenience, but the existence is of great value as a standard for comparison with the more conventional type.

A criterion for use in the optimization of a closed loop system may be formulated as follows:

Let $z(t)$ be the desired system output, and let $c(t)$ be the actual system output. Then any functional of $c(t)$ and $z(t)$ is some kind of measure of how well the system operates. Usually, a measure of system performance is some quantity which depends on the error

$$z(t) - c(t)$$

and which is a minimum when the error is zero, and becomes larger when error is increased.

The most important criterion in use is the mean square error criterion

$$\int [z(t) - c(t)]^2 \, dt$$

which is used mainly because it permits the development of analytic methods for synthesizing systems with random inputs. For many situations, other criteria are more suitable but suffer from lack of mathematical development. This criterion also has a physical interpretation of discriminating against the occurrence of large errors, i.e., the optimization criterion of minimum mean square ensemble system error weights large errors more heavily than small errors. It reduces the likelihood of large errors but leaves the system relatively sensitive to small errors.

However, other criterion might be more important in some cases. For example, it might be more appropriate to maximize the probability that the error be less than some prescribed tolerance,

$$\text{Prob} \quad \{|Z(t) - c(t)| < K\}$$

i.e., we require a system which minimizes the specified probability. All errors greater than a certain threshold are equally bad, while small errors are tolerated.

$$\text{Prob} \quad \{|Z(t) = c(t)| \; s(\tau), \; \tau \leq t\}$$

i.e. we require the system whose output has the largest conditional probability, using all the past history of the signal, of being the correct value. But this requires continuous conditional probabilities; assumes all errors larger than a certain value are equally bad; and requires a complete statistical knowledge of inputs, often not available.

XI. OPTIMAL FILTERING AND PREDICTION

Conventional frequency filters are intended to separate

two classes of signals whose spectra do not overlap. Specification of these filters does not depend on the statistical properties of the signals.

However, communication and control systems often must perform the task of separating as well as possible a desired signal input from an extraneous signal input such as random noise, whose spectrum greatly overlaps that of the desired signal. Optimum filters (relative to the way they are specified) are a class of filters designed to perform this separation.

Assume we have a corrupted signal y(t) which is the sum of a desired signal s(t) and unwanted noise n(t)

$$y(t) = s(t) + n(t) .$$

Smoothing is the removal of the unwanted random roughness in the data. In some cases n(t) has higher frequency components than s(t), and removal of n(t) actually amounts to smoothing the graph of the signal.

Predicting is the forecasting of a future value of the desired input signal.

Smoothing and predicting can be combined together, as well as with and without differentiation.

Two fundamental principles of smoothing and predicting are:

1. No separation of signal s(t) from s(t) + n(t) is possible unless s(t) and n(t) have distinguishing properties.

2. No prediction of s(t) into the future can be made unless it has known property which relates its past and future, at least in some statistical sense.

The term "filter" or "memory function" used hereafter is to be taken in a very broad context. Although mathematically it may be termed an "operator," physically in this sense it may be a suitable electric circuit or a more complicated piece of equipment, such as an automatic feedback control loop; or a memory store containing a given set of

transmitted messages for comparison with a garbled message received; or various electronic circuits, or even a missile itself.

Consider the general system

A smoothing filter is designed to extract as well as possible a desired signal s(t) from the mixture y(t) of signal and noise. Here T = 0, and the system is also called a "duplicator."

A predicting filter is designed to yield a future value s(t + T), $T \geq 0$ of the signal, where the signal s(t) may or may not ($n(t) \neq 0$) be mixed with noise in the present.

The over-all problem of designing systems to perform smoothing and predicting can be considered in two parts:

1. Determine by some criterion what the best smoothing and predicting is, and what the optimum performance operator is.

2. Realization of the desired performance operator in a workable device.

Until the work of Wiener became known, the design of linear systems depended on a combination of cut-and-try procedures and analytical methods for choosing, in some optimum fashion, the free parameters of a system of given form. Norbert Wiener (1942) solved the problem of optimum prediction and filter under these four assumptions:

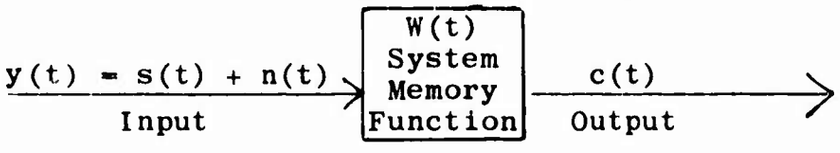
1. The system is a fixed parameter-linear device (since extended to a time-varying parameter linear device).

2. The system has infinite memory, i.e., operates on all past history of the signal and noise (since extended to include the cases of finite duration sampling time).

3. The input time series (both signal and noise) are ergodic stationary random processes (since extended to the case of nonstationary nonergodic random process input).

4. The system is optimum in the sense of minimization of least squares (this has <u>not</u> been generalized in a practical manner).

The Wiener method of system optimization specifies mathematically that memory function which makes output, $c(t)$, the "best" approximation to a translation of the message input $s(t + \tau)$ where τ is some real number.



The error in the approximation, $\epsilon(t)$, is defined as

$$\epsilon(t) = s(t) - c(t)$$

and the approximations will be assumed "best" when the mean square of $\epsilon(t)$ is minimum, i.e.,

$$\overline{\epsilon}^2 = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} [\epsilon(t)]^2 \, dt$$

The mathematical application of the Wiener method is as follows:

1. Express the mean squared error in terms of statistical properties of the message and noise signals.

2. Minimize the mean squared error by use of the calculus of variations. In the process of minimization, it will be found that the optimum memory function must satisfy a Wiener-Hopf type integral equation.

3. The desired optimum memory function will be expressed in terms of the solution to the integral equation and characterized by transfer function relating the output to the input.

(Memory function is to be regarded in the broad context as a component, subsystem, or even a system, depending on the problem.)

Some important limitations on the Wiener method which must be kept in mind are:

1. All physical systems to which this method is applied are linear.

2. In many cases, it is impossible to realize practically the memory function derived analytically by the Wiener method. The design is feasible provided the optimum process does not require a circuit to distinguish between positive and negative frequencies, and does not require it to have a negative memory. (A negative memory would imply that a signal could be processed before it had been received.

3. The auto- and cross-correlation functions and spectral densities functions for the noise must exist.

A relevant question is "Why bother with optimum theory, when the devices may turn out to be physically unrealizable?" Two reasons exist.

1. It is desirable to have a standard of reference representing the maximum attainable performance that can be expected for an optimum system. By considering the theoretical performance capabilities of optimum systems, one can often show that a simple, practical system under study may differ so little from optimum as to make further refinement unnecessary.

2. Even though optimum filters may not of themselves be easily instrumented, they can often be approximated by practical devices.

The Wiener method yields an "absolute" optimum system memory function, whereas the so-called Phillip's method derives a "relative" optimum system memory function.

This relative optimum is found by assuming a basic structure for a system and optimizing with respect to its controllable parameter.

Note that the matter of physical reality, a severe limitation of the Wiener method, does not enter and no consideration is given whether or not the optimum system could be practically realized. Since the basic structure is fixed prior to optimization, the question does not arise. An optimum memory function is then derived as follows:

1. Establish an expression for the mean-squared error of the system;

2. Minimize this error with respect to controllable parameters.

For example, the realizable linear-over-quadratic transfer function

$$\frac{as + b}{cs^2 + ds + e}$$

as a beam riding computer filter for a guided missile can be shown to approximate the results of the theoretical optimum filter for the case of "white" noise (all frequencies represented) and game theory acceleration spectra (discussed later).

It is unfortunate that guidance filtering which is optimal from the standpoint of minimizing the maximum mean square miss distance (or beam riding error) demands infinite mean square missile acceleration. It seems feasible to take into account the actual physical limiting of missile acceleration by seeking the filter which minimizes the mean square miss (or beam riding error) subject to a constraint on the allowed mean square missile acceleration.

However, by separating the guidance and control functions

in the missile, we can include the effects of target maneuver in the optimization of the beam riding computer with drag-induced slowdown as a constraint, and include the effects of noise in the optimization of the autopilot with acceleration limiting as the constraint occurring there.

Eventually, we define the optimum beam riding computer to be that one which minimizes total beam riding errors due to both noise and target maneuvers while restricting the induced drag due to noise to some acceptable value, i.e., an acceptable value of induced drag is that value which does not cause the missile to slow down in level flight at the maximum intended range.

The near-optimum transfer function is no longer linear-over-quadratic, but becomes linear over-quartic, due to the constraints.

As a side note, we should mention that while optimization of filters concerned with smoothing and predicting data has, on the whole, been done with respect to the mean square error criterion, the filters are sometimes complicated to compute, making them quite often unsuited for real time solutions. For such applications, it may be necessary to optimize the filters from the primary viewpoint of ease of computation, with which the final estimates of the output are obtained, and optimization in the mean square value sense is obtained as a secondary consideration in order to provide some control of the mean square output error. This approach would have the greatest appeal for problems involving real time filtering where computing time and complexity are primary considerations, and noise reduction is a secondary consideration.

A remark on the use of nonlinear filters can be made here. If the noise and signal at the input both possess Gaussian distributions in amplitude, the linear filter is the optimum filter, and no improvement in filtering can be realized by going to a nonlinear device. But with more general input signals, the same situation does not hold. In many cases the mean square error can be further reduced by the addition of nonlinear filters.

In concluding this section, we can state a few facts about detection of signals in noise. Detection is the process of determining whether a signal is present or not. When detecting a signal, the detector is either right or wrong, but in the theory of prediction, it is unlikely that one is ever

exactly right, but there are all degrees of wrongness. In detection, alternative optimization criteria to minimization in a least squares sense has been used. For example:

1. Maximization of signal-to-noise ratio at a specified instant of time;

2. Maximization of the absolute magnitude of the differences between signal and noise over all time;

3. In one optimum threshold (i.e., weak signal) detection system, optimality is achieved by minimizing the average cost of decision;

4. Another detection system is "best" which in the long run will hold fixed the false alarm probability and will minimize the probability of missing the signal.

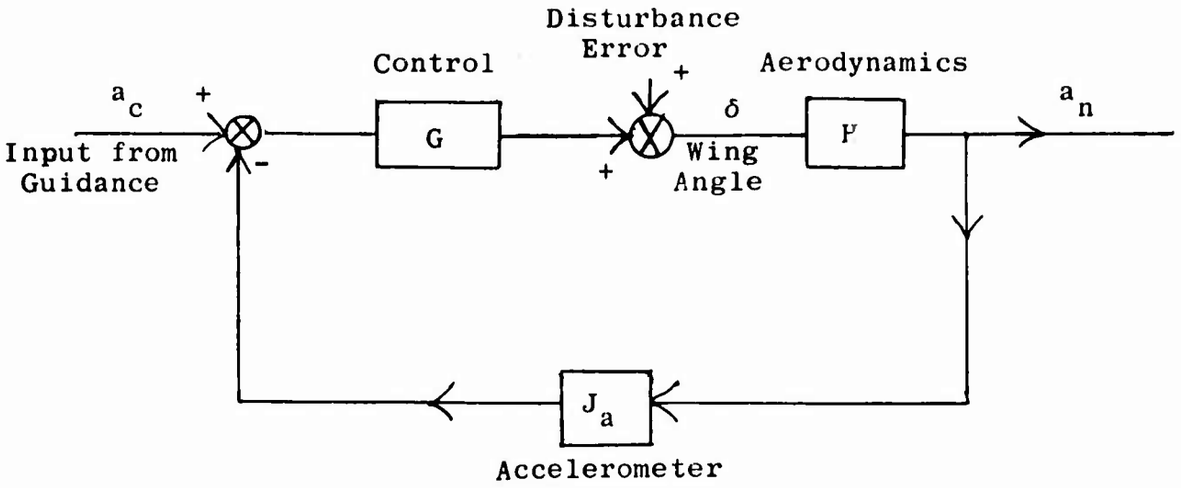
XII. WIENER THEORY APPLIED TO AUTOPILOT DESIGN

The Wiener optimum linear filter formulation is applied to the autopilot closed loop transfer function with the disturbance spectrum entering at the wing. [Actually the noise in the autopilot enters the loop at various places.]

A statistically optimum autopilot may be defined as that autopilot which minimizes the mean-square error between acceleration command and the missile acceleration response attained.

For a realistic approach, saturation of various elements in the system must be included. Because of the nonlinear nature of these saturating elements, general procedures for the optimization of filters involving them have not been developed. Usually the nonlinear limits are replaced by a linear system by constraining the saturating quantities to their mean square values.

It is desired to find the transfer function G which minimizes the mean square $\overline{\epsilon^2}$, subject to a constraint of limited mean square wing rate $\overline{\dot{\delta}^2}$.



XIII. GAME THEORY AND OPTIMAL FILTERING

The theory of games, or game theory, is the particular branch of applied mathematics which deals with the rational analysis of competitive and cooperative systems, and the determination of optimal courses of action for the participants. Situations to which game theory can be applied occur in operations analysis, strategic and tactical planning, whenever there is conflict of interest arising from the actions of an opponent. ("Opponent" can include even Nature, a fictitious player, having no known objective nor strategy in some general cases.)

The elementary concepts for the simplest, two-person game are that which one side gains the other side loses, that the opponents simultaneously choose a course of action (called a strategy), and that the outcome of the game (the payment of one side to the other) is determined by this dual choice of strategies. The outcome is not completely determined by either side alone, it is determined by the combined decisions of the two opposed players.

One of several possible criteria is the minimax principle -- each player should employ only optimal strategies which minimize his maximum loss, no matter what the other player may do.

In Wiener's work, the statistical properties of the noise and signal are assumed fixed, and given in advance of the problem. A guidance filter can be optimized then to include the effects of target random maneuvers, presumably typified by maneuvers over which the enemy himself has little or no control [e.g., wind gusts, noise in his system, etc.]. Such a guidance filter or beam riding computer may or may not be near optimum if the target maneuvers in a deliberate manner. Essentially, the game theory approach assumes an intelligent target, one able to assume the most evasive maneuvers within its capabilities. The goal is to design the best possible system against this crafty target, within certain boundary conditions and constraints.

In game theory optimization, allowance is made for the fact that the enemy who is producing the signal may prefer not to be followed, and may attempt to keep the mean-square error as large as possible. The spectral density of his signal is thus no longer fixed, but is the strategy of one of the participants of the game.

The strategy of the other participant, the filter designer, is specified by the transfer function of the filter. The game payoff is considered to be the mean-square error or miss.

The filter designer must design the best filter in view of the worst choice of signal spectral density by the enemy (as caused by his maneuver).

The enemy, who is the signal producer, must generate a signal which has statistical characteristics that result in the greatest possible miss despite the best efforts of the filter designer.

If these two considerations are mutually compatible, there is a game theory solution, i.e., there is a filter transfer function which gives the smallest mean-square error for the worst possible signal.

The game theory solution is essentially the intersection of two functional equations: one giving the optimum filter transfer function for fixed signal spectral density; the other the optimum signal spectral density for fixed filter function. This solution has the property that the error obtained with it when the target signal strategy is optimum (in the sense of

maximizing the miss) is smaller than the error obtained for any other filter transfer function when the signal strategy is optimum for that transfer function.

XIV. WHY OPTIMIZATION TECHNIQUES HAVE NOT BEEN USED MORE FREQUENTLY

Three conditions have conspired to keep optimization methods from widespread application in the past.

1. Complicated mathematics, for example, in optimizing a filter by Wiener's method, and actual nonlinear constraints.

2. Equipment required is prohibitive in cost and/or bulk.

3. Advantages accruing from successive refinements become progressively smaller. Quite often a comparatively simple filter, say roughly approximating the Wiener design, in practice gives most of the advantage to be expected from the true Wiener design.

XV. RELATION TO ADAPTIVE SYSTEMS

Designing control systems where (a) little significant information is known about the process to be controlled, (b) the properties of the process vary over an extraordinarily large range, and (c) the characteristics of the system input signals change markedly with time, may require systems in which the compensation is automatically adjusted to offset these adverse effects. Such systems are called "adaptive systems."

Optimum performance from operating systems is of increasing importance as competition in the real world becomes more severe. Adaptive control is a method of automatic control aimed at obtaining optimum system performance even when there exists incomplete or inexact analytical or analog models of the process that is being controlled.

Adaptive or self-optimizing systems optimize control by

the use of an automatic unit which searches out and holds the best performance from a controlled system, in spite of any reasonable change of the output level or environmental operating conditions. In this sense, the adaptive system "learns" to improve its performance, based upon experience, thereby adapting itself to the circumstances it finds.

In order for an optimizing action to be possible, the system to be controlled must have a performance characteristic which shows an optimum point as one or more of the inputs vary. A suitable deviation signal that represents the departure of the operation from the optimum condition is generated and utilized as a basis for making corrections to the input (if controllable) or to the adaptive system parameters (e.g., variable damping if noise is the input), so that the deviation diminishes and the operating point approaches the optimum point again.

An application of adaptive systems is made in target tracking radars, which adjust the over-all transfer function of the tracking loop according to target behavior and system noise. Target acceleration or maneuver is "recognized" and the system parameters vary to convert the function of the loop from data smoothing to tight and fast follow-up. When target maneuver ceases, system returns to its filtering mode. Thus, the parameters are adjusted to balance the low frequency error (lag due to maneuver) and the high frequency error (resporse due to noise).

XVI. MATHEMATICAL PROGRAMMING

Linear programming, nonlinear programming and dynamic programming are three of a set of techniques called "Mathematical Programming" which involve the programming of interdependent activities. These techniques are referred to as "programming" in order to emphasize that planning, as distinguished from operations or execution of plans, is the area of primary interest. In brief, programming is concerned with the problem of planning a complex of interdependent activities in the best possible way.

The general programming problem is to maximize or minimize an objective function $\phi(x)$, which is some over-all measure such as cost, or profit, or value, or quality, or efficiency, etc. subject to constraints $g_1(x) \geq 0$, $g_2(x) \geq 0$, ..., and $x_1 \geq 0$, $x_2 \geq 0$, ...

The usual analytic methods of solving extremization problems in the presence of constraint equations (solving the constraint equations, substituting in the objective function, and differentiating, or else by Lagrange multiplier techniques) do not take into account the inequality constraints which characterize mathematical programming.

XVII. LINEAR PROGRAMMING

Linear programming is a relatively new mathematical technique to handle problems with the following characteristics.

1. There is usually a large but finite number of non-negative variables.

2. The variables are subject to a finite number of constraints or boundary conditions usually in the form of linear inequalities and/or equations which limit the variables range, and are accurately known.

3. Under these constraints, some objective function is to be maximized or minimized.

Both the objective function to be maximized and the restrictions on each variable (equalities or inequalities) are linear in the variables.

We say a solution is feasible if it satisfies the constraints, and optimal if it also achieves a maximum. The problem consists in determining, out of the infinite number of feasible solutions, a unique (if possible) solution which is optimal.

The resulting solution will then provide the best possible planning of operations under the specified restrictions.

Problems with these characteristics crop up in transportation fields (relating sources and destinations of supplies), production bottlenecks (efficient allocation of limited resources) problems of scheduling and timing, contract awards, personnel assignment, etc.

Linear programming problems have been attacked in several ways, the most prominent method being the Simplex Algorithm, which is a very powerful computational technique which can efficiently solve large systems containing hundreds of equations.

However, limitations and disadvantages to linear programming exist. Among them are:

1. The linearity restriction. For example, linear costs do not penalize large values of the variable. [This has been alleviated to some ex tent by the generalization to "Quadratic Programming," in which the linear objective function to be extremized can be replaced by a quadratic form, with linear constraints.]

2. Optimal solutions are not obtained in analytic form. Changes in the mathematical model require recalculation.

3. Error analysis is difficult.

4. No provision is made for relationships involving uncertainty due to random fluctuations or errors in determination, for example, sales forecast in the form of a probability distribution cannot be handled.

5. The expression of realistic objectives and constraints in measurable terms.

6. The determination of suitable numerical values for coefficients.

7. The computational labor required to execute numerically large-scale linear programming problems.

XVIII. DYNAMIC PROGRAMMING

Among programming problems, there are some in which time plays an essential role and in which the sequence of decisions is vital. These are termed "Dynamic Programming" problems, and Dynamic Programming, the functional equation technique of a new mathematical discipline, can be used in the formulation and solution of optimization problems, including those in which the process need not necessarily change with time. Furthermore, the process may be stochastic, i.e., the outcome is not determined but is predictable by means of a probability distribution.

In a typical application, a sequence of decisions is sought which in some sense optimizes the behavior of a sytem. In these sequences of operations the outcomes of the preceding operations may be used to guide the course of future operations. As the number of decisions increases, and the discrete length of the decision interval decreases without limit, continuous solutions are produced which are equivalent to those furnished by the classical calculus of variations. Thus, Dynamic Programming is properly an extension of the calculus of variations, but of much wider scope and versatility.

XIX. QUEUEING THEORY

In many operations there is a lack of timing between arrival, at some point in the operation, of a sequence of units, and the subsequent disposal of these units, so that a waiting line or queue is formed of newly arrived units awaiting disposal. For example, the units might be aircraft "stacked" over an airport, traffic tie-up, ships in a harbor awaiting docking, customers in a cafeteria, etc. Queueing Theory or Waiting Line Theory is the specialized method for analysis of these situations.

A central problem of waiting line theory is the relationship between the mean length of the waiting line and the degree of randomness of arrival and disposal, such a line arising whenever the mean arrival rate exceeds the mean service rate. On this problem can be based estimates of the optimum capacity of the service facilities when one balances the cost of letting the unit wait in line against the cost of increasing the service rate.

XX. THE CONVERSE OF OPTIMAL

The converse of "optimal" i.e., the worst of the worst has been dubbed "pessimal" by J L. Vanderslice (APL). However, this term is not in common usage in the technical journals.

BIBLIOGRAPHY

(Items are listed in their approximate occurrence in the text.)

1. Saaty, T., Mathematical Methods of Operations Research, McGraw-Hill, 1958, see Chapter 5, "Optimization," in particular.

2. Modern Mathematical Methods and Models, Volume I: Multicomponent Methods, by the Dartmouth College Writing Group, see Chapter 4, "Optimization Problems."

3. Hitch, C., "Suboptimization in Operations Research," Journ. Op. Res. Soc. Am., Vol. 1, No. 3, May 1953, pp.87.

4. Oakley, C. O., "End-Point Maxima and Minima," Am. Math. Monthly, Vol. 54, p. 407.

5. Monall, J. D., "One-Sided Maxima and Minima," Am. Math. Monthly, Vol. 55, p. 311.

6. Fox, Charles, An Introduction to the Calculus of Variations, Oxford University Press, 1950.

7. Forsythe, G. E., "Computing Constrained Minima with Lagrange Multipliers," Journ. Soc. Ind. Appl. Math., Vol. 3, No. 4, December 1955.

8. "Report of a Symposium on Modern Techniques for Extremum Problems," Operations Research, Vol. 5, No. 2, April 1957, p. 244.

9. Lewis, E. V., "Optimum Fullness for Dead Weight Cargo Ships in Moderate Weather Service," Journ. Ship Research, November 1957, p. 7.

10. Plunkett, R., "The Calculation of Optimum Concentrated Damping for Continuous Systems," Journ. Applied Mechanics, Vol. 25, No. 2, June 1958, p. 219.

11. Brooks, S., "A Discussion of Random Methods for Seeking Maxima," Op. Res., Vol. 6, No. 2, March-April 1958, p. 244.

12. Best, G. C., "A Minimum Problem Solved by Mesh Methods," MTAC, Vol. 8, 1954, p. 11.

13. Vasu, G., "Experiments with Optimalization Controls Applied to Rapid Control of Engine Pressures with High-Amplitude Noise Signals," Trans. ASME, Vol. 79, No. 3, April 1957, p. 481.

14. Herget, P. and Clemence, G., "Optimum Interval Punched Card Tables," MTAC, Vol. 1, 1944, p. 173. See also a note by J C. P. Miller on p. 334.

15. Saddler, D H., "Maximum Interval Tables," MTAC, Vol. 4, 1950, p. 129.

16. Blackman, N., "Minimum-Cost Encoding of Information," IRE Trans. Inform. Theory, PGIT-3, March 1954.

17. Huffman, D., "A Method for the Construction of Minimum-Redundancy Codes," Proc. IRE, September 1952.

18. Wilkinson, J. H., "An Assessment of the System of Optimum Coding used on the ACE at the National Physical Laboratory," Phil. Trans. Roy. Soc. (A), 20 October 1955, p. 253.

19. Gordon, B., "An Optimizing Program for the IBM 650," Journ. Assoc. Comp. Mach., Vol. 3, No. 1, January 1956, p. 3.

20. Breakwell, J. V. "The Optimization of Trajectories," Journ. Soc. Ind. Appl. Math., Vol. 7, No. 2, June 1959, p. 215.

21. Newton, R., "On the Optimum Trajectory of a Rocket," Journ. Franklin Institute, Vol. 266, No. 3, September 1958.

22. Koopman, B. O., "The Optimum Distribution of Effort," Journ. Op. Res. Soc. Am., Vol. 1, 1953, p. 52.

23. Miehle, W., "Numerical Solution of the Problem of Optimum Distribution of Effort," Journ. Op. Res. Soc. Am., Vol. 2, 1954, p. 433. Also correction, p. 219, Vol. 3, 1955.

24. Koopman, B. O., "The Theory of Search," published in Operations Research Journal in three parts.

Part I. "Kinematic Bases," Vol. 4, No. 3, June 1956.

II. "Target Detection," Vol. 4, No. 5, October 1956.

III. "The Optimum Distribution of Searching Effort," Vol. 5, No. 5, October 1957.

25. Graham, D. and Lathrop, R., "The Synthesis of Optimum Transient Response: Criteria and Standard Forms," Trans. AIEE, Vol. 72, Part II, November 1953, p. 272.

26. Schultz, W and Rideout, V., "A General Criterion for Servo Performance," Proc. Nat. Elect. Conf., Vol. 13, 1957, p. 459.

27. Oldenburger, R., "Optimum Nonlinear Control," Trans. ASME Vol. 79, No. 3, April 1957, p. 527.

28. Davenport, W. and Root, W., An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, 1958.

29. Bendat, J., Principles and Applications of Random Noise Theory, J. Wiley, 1958.

30. Newton, G., Gould, L., and Kaiser, J., Analytical Design of Linear Feedback Controls, J. Wiley, 1957.

31. James, H., Nichols, N., and Phillips, R., Theory of Servomechanism, Vol. 25, MIT Radiation Lab Series, McGraw-Hill, 1947.

32. Ragazzini, J., and Zadeh, L., "Probability Criterion for the Design of Servomechanisms," Journ. Appl. Physics, Vol. 20, February 1949, p. 141

33. Shinbrot, M., and Carpenter, G., An Analysis of the Optimization of a Beam-Rider Missile System, NACA, Tech. Note 4145, 1958.

34. Bode, H. and Shannon, C., "A Simplified Derivation of Linear Least-Squares Smoothing and Prediction Theory," Proc. IRE, Vol. 38, April 1950, p. 417.

35. Darlington, S., "Linear Least Squares Smoothing and Predicting, with Applications," Bell System Technical Journal, September 1958, Vol. 37, pp. 1221-1294.

36. Stromer, P. R., "Adaptive or Self-Optimizing Control Systems - A Bibliography," IRE Trans. on Automatic Control, Vol. AC-4, No. 1, May 1959, p. 65.

37. Aseltine, J., Mancini, A., and Sarture, C., "A Survey of Adaptive Control Systems," IRE Trans. on Automatic Control, PGAC-6, December 1958, p. 102.

38. Munson, J., and Rubin, A., "Optimization by Random Search on the Analog Computer," IRE Trans. on Electronic Computer, Vol. EC-8, No. 2, June 1959.

39. Vajda, S., The Theory of Games and Linear Programming, Methuen, 1956.

40. Bellman, R., Dynamic Programming, Princeton U. Press, 1957.

41. Morse, P. M., Queues, Inventories, and Maintenance, ORSA Publications No. 1, J. Wiley, 1958.

42. Saaty, T., "Resume' of Useful Formulas in Queueing Theory," Operations Research, Vol. 5, No. 2, April 1957, p. 161.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

OPTIMIZATION IN NOISE

by

J. E. Hanson

I. LINEAR FILTERS, THEIR TRANSFER FUNCTIONS AND WEIGHTING FUNCTIONS*

Let x(t) and y(t) be the input and output, respectively, of a filter. By this we understand that there exists a linear differential equation with constant coefficients connecting x(t) and y(t). We shall temporarily assume this equation to have the form

$$(1.1)\qquad b_n y^{(n)} + b_{n-1} y^{(n-1)} + \ldots + b_1 \dot{y} + b_o y = a_{n-1} x^{(n-1)} + \ldots + a_1 \dot{x} + a_o x$$

where

$$(1.2)\qquad b_n \neq 0.$$

Replacing the operator $\frac{d}{dt}$ by the letter S, (some people use D, others P), we can formally solve for y, thus:

$$(1.3)\qquad y = \frac{a_{n-1} S^{n-1} + \ldots + a_1 S + a_o}{b_n S^n + b_{n-1} S^{n-1} + \ldots + b_1 S + b_o}\, x = F(S)\, x.$$

The rational function of S, F(S), is called the transfer function of the filter. In elementary differential

*Some of the discussion of this and subsequent sections is simplified, and not completely rigorous mathematically. The purpose of this paper is to acquaint the reader with useful concepts and techniques, which is oftentimes at odds with the objectives of completeness and rigor.

equations it is shown that if one solves for the roots of $b_n S^n + b_{n-1} S^{n-1} + \ldots + b_1 S + b_o = 0$, and if the roots all have negative real parts, then the equation (1.1) is stable, i.e., if $x(t) = 0$, then $y(t)$ will approach zero no matter what the initial values of $y, \dot{y} \ldots, y^{(n-1)}$. Accordingly, if the denominator of $F(S)$ (extreme caution must be exercised in cancelling common factors of numerator and denominator) has the ascribed property, we shall say the filter is stable or realizable. (The term "realizable" is found in the literature but its use in this connection is objectionable.)

Let the filter be stable, and be of the form given by (1.1). The following statements are then valid:

A. There exists a unique function $W(\tau)$, called the weighting function of the filter, defined for $\tau \geq 0$ such that, for any sufficiently well behaved input, we have in steady state

$$y(t) = \int_0^\infty W(\tau)\, x\,(t-\tau)\, d\tau, \text{ assuming } x(U)$$

defined for $-\infty < U \leq t$.

B. $F(S) = \int_0^\infty e^{-S\tau}\, W(\tau)\, d\tau$, for complex S with sufficiently large real parts. (ReS = 0 is large enough in all cases).

C. $F(\bar{S}) = \overline{F(S)}$, $F(S)$ real for real S.

D. If $x(t) = A \cos \omega t$*, then in steady state

$$y(t) = A\left|F(i\omega)\right| \cos(\omega t + \phi), \text{ where}$$

$$\tan \phi = \frac{\text{Im } F(i\omega)}{\text{Re}F(i\omega)} .$$

Justifications---We shall conclude this section with justifications of A, B, C, D and a short discussion of the significance of them.

---

*A is a constant.

Proof of A

For any sufficiently well behaved function $W(\tau)$, we have, setting $y(t) = \int_0^\infty W(\tau)\ x\ (t-\tau)\ d\tau$.

$$(1.4) \qquad b_o y = \int_0^\infty b_o\ W(\tau)\ x\ (t-\tau)\ d\tau\ .$$

$$(1.5) \qquad b_1 \dot{y} = \int_0^\infty b_1\ W(\tau)\ \dot{x}\ (t-\tau)\ d\tau = b_1\ W(0)\ x\ (t)$$

$$+ \int_0^\infty b_1\ \dot{W}(\tau)\ x\ (t-\tau)\ d\tau$$

using integration by parts. Using successive integration by parts, for $k \leq n$ we have

$$(1.6) \qquad b_k\ y^{(k)} = \int_0^\infty b_k\ W(\tau)\ x^{(k)}\ (t-\tau)\ d\tau = b_k\ W(0)\ x^{(k-1)}(t)$$

$$+ b_k\ \dot{W}(0)\ x^{(k-2)}\ (t) + \ldots + b_k W^{(k-1)}(0)\ x(t)$$

$$+ \int_0^\infty b_k\ W^{(k)}(\tau)\ x\ (t-\tau)\ d\tau\ .$$

Adding up the above n+1 equations, there results

$$(1.7) \qquad b_n\ y^{(n)} + b_{n-1}\ y^{(n-1)} + \ldots + b_1\ \dot{y} + b_o y$$

$$= \int_0^{\infty} [b_0 W(\tau) + b_1 \dot{W}(\tau) + \ldots + b_n W^{(n)}(\tau)] \times (t-\tau)\, d\tau$$

$$+ b_n W(0)\, x^{(n-1)} + [b_n \dot{W}(0) + b_{n-1} W(0)]\, x^{(n-2)}$$

$$+ \ldots + [b_n W^{(n-2)}(0) + \ldots + b_3 \dot{W}(0) + b_2 W(0)]\, \dot{x}$$

$$+ [b_n W^{(n-1)}(0) + \ldots + b_2 \dot{W}(0) + b_1 W(0)]\, x \, .$$

Note that if we can satisfy

$$b_n W^{(n)}(\tau) + \ldots + b_1 \dot{W}(\tau) + b_0 W(\tau) = 0, \text{ for } \tau \geq 0, \tag{1.8}$$

and

$$b_n W(0) = a_{n-1} \tag{1.9}$$

$$b_n \dot{W}(0) + b_{n-1} W(0) = a_{n-2}$$

$$\vdots$$

$$b_n W^{(n-2)}(0) + \ldots + b_3 \dot{W}(0) + b_2 W(0) = a_1$$

$$b_n W^{(n-1)}(0) + \ldots + b_2 \dot{W}(0) + b_1 W(0) = a_0$$

all simultaneously, then the integral in (1.7) vanishes identically, and y(t) is then a particular solution to (1.1). The general solution, from elementary differential equations can then be expressed as the sum of y(t) and the general

solution to the reduced differential equation, the latter approaching zero as $t \rightarrow \infty$. Thus, if we wait long enough after the initial instant, the output will be approximated as closely as we please by y(t). This is what is meant by the steady state solution. If we have not waited very long, the difference between the output and the steady state solution is called a transient.

The equations (19) can be solved for W(0), ..., $W^{(n-1)}(0)$ uniquely since $b_n \neq 0$. Thus $W(\tau)$ is a solution to our problem if it is a solution of the reduced differential equation with specified initial conditions. That such a solution exists and is unique follows from differential equation theory. We have now found a function $W(\tau)$ such that

$$(1.10) \qquad y(t) = \int_0^\infty W(\tau) \times (t-\tau)\, d\tau$$

is the steady state output of the filter. It can also be shown that there is only one such function $W(\tau)$ with this property. (The proof depends on complex variable theory, and is omitted).

Note that $W(\tau)$ will decay exponentially to zero as $\tau \rightarrow \infty$.

If we allow the numerator of F(S) to have degree equal to or larger than that of the denominator, the previous analysis breaks down. The concept of weighting function can be extended even so, however, but not without introducing the Dirac delta function. The $W(\tau)$'s are no longer nice functions in the usual sense, and their manipulations are fraught with hazards. Even a competent person in filter theory must occasionally exercise caution, although he is generally quite familiar with delta functions and their admissable manipulations.

Transfer functions whose denominators have roots with positive real parts are seldom purposely used in missile work as they represent unstable filters. However, quite often it is useful to consider roots with real parts equal to zero. (For example, $\frac{1}{S}$ would formally represent a pure integrator.) If one pushes the root slightly to the left in the

complex plane (say $\frac{1}{S+\epsilon}$), proceeds with the analysis, and then finally lets $\epsilon \to 0$, one can justify many of the manipulations executed by a competent servo man, who generally neglects this necessary logical step. He knows by experience what he can and cannot do with these transfer functions.

One can also justify the delta function manipulations for the case mentioned previously by adding a few small higher order terms to the denominator to reduce the transfer function to the case we have considered, and then at the end letting these terms approach zero.

In general, we shall try to avoid these somewhat pathological transfer functions, leaving the pursuit of them to the interested reader.

Proof of B

Applying integration by parts liberally, we can write

$$(1.11) \quad \int_0^\infty b_o\, e^{-S\tau}\, W(\tau)\, d\tau = b_o \int_0^\infty e^{-S\tau}\, W(\tau)\, d\tau$$

$$\int_0^\infty b_1\, e^{-S\tau}\, \dot{W}(\tau)\, d\tau = b_1\, S \int_0^\infty e^{-S\tau}\, W(\tau)\, d\tau - b_1\, W(0)$$

$$\int_0^\infty b_2\, e^{-S\tau}\, \ddot{W}(\tau)\, d\tau = b_2\, S^2 \int_0^\infty e^{-S\tau} W(\tau)\, d\tau - b_2\, \dot{W}(0)$$

$$-b_2\, S\, W(0)$$

$$\vdots$$

$$\int_0^\infty b_n\, e^{-S\tau}\, W^{(n)}(\tau)\, d\tau = b_n\, S^n \int_0^\infty e^{-S\tau} W(\tau)\, d\tau - b_n$$

$$W^{(n-1)}(0) - b_n\, SW^{(n-2)}(0) - \ldots - b_n S^{n-1} W(0) \quad .$$

Adding all of the above, using (1.8) and (1.9), we have

$$0 = (b_o + b_1 S + \ldots + b_n S^n \int_o^\infty e^{-S\tau} W(\tau)\, d\tau \tag{1.12}$$
$$- (a_{n-1} S^{n-1} + \ldots + a_1 S + a_o).$$

Solving for the integral yields the desired result. The above operations are legitimate as long as the real part of S is greater than the real part of every root of the denominator of F(S).

Proof of C

This is obvious, and the proof is omitted.

Proof of D

Since

$$A \cos \omega t = \frac{A}{2} (e^{i\omega t} + e^{-i\omega t}), \tag{1.13}$$

we have, from (1.10) and statements B and C,

$$y(t) = \frac{A}{2} \int_o^\infty W(\tau) [e^{i\omega(t-\tau)} + e^{-i\omega(t-\tau)}]\, d\tau \tag{1.14}$$

$$= \frac{A}{2} e^{i\omega t} \int_o^\infty W(\tau) e^{-i\omega\tau}\, d\tau$$

$$+ \frac{A}{2} e^{-i\omega t} \int_o^\infty W(\tau) e^{i\omega t}\, d\tau$$

$$= \frac{A}{2} [e^{i\omega t} F(i\omega) + e^{-i\omega t} F(-i\omega)]$$

$$= \frac{A}{2} [\cos \omega t + i \sin \omega t] [\mathrm{Re}F(i\omega) + i\, \mathrm{Im}\, F(i\omega)]$$

$$+ \frac{A}{2} [\cos \omega t - i \sin \omega t] [\mathrm{Re}F(i\omega) - i \, \mathrm{Im} \, F(i\omega)]$$

$$= A [\mathrm{Re} \, F(i\omega) \cos \omega t - \mathrm{Im} \, F(i\omega) \sin \omega t].$$

The result follows by setting

$$(1.15) \qquad \cos \phi = \frac{\mathrm{Re} \, F(i\omega)}{|F(i\omega)|}$$

and

$$(1.16) \qquad \sin \phi = \frac{\mathrm{Im} \, F(i\omega)}{|F(i\omega)|} \quad .$$

Discussion---Statement D is a key to the usefulness of the concept of a transfer function, and to why the word "filter" is used. It tells us what happens to the output of a filter in steady state when the input is a sine wave. $|F(i\omega)|$ may be large for some ω's and small for others. Since the ratio of the amplitude of output to input is just $|F(i\omega)|$, we see that when $|F(i\omega)|$ is small, the filter "filters out" that frequency, i.e., it greatly attenuates its amplitude.

When one uses transfer functions, one is commonly said to be "working in the frequency domain", mainly because of the above paragraph, i.e., because of the natural connection between behavior of filters in the presence of sine wave inputs and pure imaginary values of S. When one analyzes filters by use of their weighting functions alone, one is said to be "working in the time domain". The conversion from one domain to the other is seen by statement B to be accomplished by a Laplace (or Fourier) Transform. (The conversion from frequency to time depends on the theory of inverse transforms, which we shall not touch upon specifically in this paper).

Exercise (a) Consider the transfer function $F(S) = \frac{1}{1+S}$.

Graph $F(i\omega)$ and $\arctan \frac{\mathrm{Im} \, F(i\omega)}{\mathrm{Re} \, F(i\omega)}$ versus ω.

Exercise (b) Let $F_1(S)$ and $F_2(S)$ be two stable transfer functions whose numerators are of lower degree than the denominators Let $x(t)$ be the input to the first, $y(t)$ its output. Let $y(t)$ be the input to the second, $z(t)$ its output. Prove that there exists a filter with transfer function $F_3(S)$ whose input is $x(t)$ and whose output is $z(t)$, and that $F_3(S) = F_1(S)$. Prove also that, for the steady state results developed in this section, it is legitimate to cancel common factors of the numerator and denominator of $F_3(S)$.

Exercise (c) If either $F_1(S)$ or $F_2(S)$ are unstable, and cancellation of factors in $F_3(S)$ reduce $F_3(S)$ to a formally stable transfer function, show that it is generally not possible to carry over the steady state results of this section to $F_3(S)$.

II. TIME SERIES, STATIONARY TIME SERIES, ERGODIC STATIONARY TIME SERIES, AUTOCORRELATION FUNCTIONS, POWER SPECTRAL DENSITIES, NOISE

By a time series we mean an ensemble (or collection, aggregate, population, class, set, etc.) of functions $x_\alpha(t)$, (α is the index which varies over the ensemble) defined for $-\infty < t < \infty$, where a probability distribution exists over the α's. Signals about which we have only statistical knowledge (such as noise, target maneuvers) are treated as being time series, since the theory of time series seems to be the only known mathematical theory whose results agree with observation, and by which one can design intelligent filters to operate with such things as inputs. A noise trace, say from missile telemetering signals of off-beam error, for example, defined for $t_o \leq t \leq t_1$, is then regarded as a section of a particular sample function of a time series.

By the autocorrelation function of a time series we mean the function $E_\alpha[x_\alpha(t)\ x_\alpha(t+\tau)]$, which is a function of t and τ. Thus, to determine the value of the function for given fixed t and τ, we compute the mathematical expectation (or mean, or average) over all α of $x_\alpha(t)\ x_\alpha(t+\tau)$.

If a time series is stationary, we think of the time series as having the same statistical properties if all functions of the ensemble are shifted to the right or left by the

same amount on the time axis. For a stationary time series, the autocorrelation function is a function of τ alone, and not of t. This simplication simplifies the mathematical theory tremendously, and fortunately, many statistical signals found in missile work can be regarded as stationary. Unless explicitly stated otherwise in the sequel, it is assumed that all time series are stationary.

The knowledge of the autocorrelation function (or equivalent functions from which it can be derived) or some approximation to it is required to do intelligent design work in many instances. With no a priori knowledge, one must resort to measurements. It is not hard to visualize the problems connected with trying to look at a lot of different samples simultaneously, being sure they are samples of the same time series, and then taking ensemble averages. Fortunately, there is a convenient crutch called ergodicity which enables one to avoid this. When looking at a sample trace of noise, one is sorely tempted to believe that one can average over time instead of over space (ensemble). So one postulates that the time series is ergodic (this term is applied only to stationary time series) and writes

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x_\alpha(t)\, x_\alpha(t+\tau)\, dt \quad \text{(for any fixed } \alpha\text{)}$$
$$= E_\alpha \left[x_\alpha(t)\, x_\alpha(t+\tau) \right] \quad \text{(for any fixed t).}$$

The autocorrelation function is then computed by the left hand side using a single noise trace. Suffice it to say that except in cases where the time series is obviously not ergodic, one rarely gets into trouble by assuming that it is. Wiener, in his classic work on stationary time series (very difficult to read) assumes ergodicity, but for most of his work, this was an unnecessary assumption. The only practical reason for assuming it is when one wishes to measure an autocorrelation function (or its equivalent). Since we will not be concerned with measurement in the sequel, we will not assume ergodicity. Henceforth, unless stated otherwise, all time series are assumed to be stationary, not necessarily ergodic, and all averages are ensemble averages.

Stationary Time Series---Let x(t) be a stationary time series, (the subscript α omitted for brevity) and let $A_x(\tau)$ denote its autocorrelation function. The $A_x(\tau)$ has the following properties:

E. $A_x(0) = \sigma^2$, the expected value of the square of $x(t)$ and hence $A_x(0) \geq 0$. (Why?)

F. $A_x(\tau) = A_x(-\tau)$ for all τ. (Replace t by $t-\tau$ in the definition of $A_x(\tau)$).

G. $|A_x(\tau)| \leq A_x(0)$ for all τ. This can be proved by the Schwarz Inequality.

We omit the proof.

The power spectral density (abbreviated by P.S.D. henceforth) denoted by $\phi_x(S)$, is defined to be

(2.1) $$\phi_x(S) = \int_{-\infty}^{\infty} e^{-S\tau} A_x(\tau)\, d\tau \quad .$$ (Laplace transform of the autocorrelation function).

For a wide class of time series, $\int_{-\infty}^{\infty} |A_x(\tau)|\, d\tau$ converges, and when this happens, $\phi_x(S)$ exists when S is a pure imaginary number, which, as one might suspect, is the case of main interest. In many cases $\phi_x(S)$ is a rational function, and can be defined over the whole complex plane (not in general by the integral, but by analytic extension). As before, we shall assume that our future manipulations are legitimate, and not be too concerned about rigor.

$\phi_x(S)$ has the following properties, which we will discuss in turn:

H. $\phi_x(S)$ is real for real and pure imaginary S.

I. $\phi_x(\bar{S}) = \overline{\phi_x(S)}$.

J. $\phi_x(S) = \phi_x(-S)$.

K. $\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_x(i\omega)\, d\omega = \sigma^2$.

L. If the time series x(t) is passed through a stable filter with transfer function F(S), the output y(t) in steady state is a time series whose P.S.D. is given by $\phi_y(S) = F(S)\, F(-S)\, \phi_x(S)$. In particular, $\phi_y(i\omega) = \left|F(i\omega)\right|^2 \phi_x(i\omega)$.

M. $\phi_x(S) \geq 0$ for pure imaginary S.

Proof of H

That $\phi_x(S)$ is real for real S follows directly from (2.1). If $S = i\omega$, $e^{-S\tau} = \cos \omega\tau - i \sin \omega\tau$. Since $\sin \omega\tau$ is an odd function, it follows from F that the imaginary part of (2.1) integrates to zero.

Proof of I

Let $S = U + iV$. Then $\phi_x(S) = \int_{-\infty}^{\infty} e^{-U\tau}(\cos V\tau - i \sin V\tau)\, A_x(\tau)\, d\tau$.

From here, I is obvious.

Proof of J

By making a change of variables, and using F,

$$\phi_x(-S) = \int_{-\infty}^{\infty} e^{S\tau} A_x(\tau)\, d\tau = -\int_{\infty}^{-\infty} e^{-SU} A_x(-U)\, dU$$

$$= \int_{-\infty}^{\infty} e^{-SU} A_x(-U)\, dU = \int_{-\infty}^{\infty} e^{-SU} A_x(U)\, dU = \phi_x(S).$$

Proof of K

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(i\omega)\, d\omega = \frac{1}{2\pi} \lim_{W\to\infty} \int_{-W}^{W} \left[\int_{-\infty}^{\infty} e^{-i\omega\tau} A_x(\tau)\, d\tau \right] d\omega$$

$$= \frac{1}{2\pi} \lim_{W\to\infty} \int_{-\infty}^{\infty} A_x(\tau) \left[\int_{-W}^{W} e^{-i\omega\tau}\, d\omega\tau \right] d\tau$$

$$= \frac{1}{\pi} \lim_{W\to\infty} \int_{-\infty}^{\infty} A_x(\tau)\, \frac{\sin W\tau}{\tau}\, d\tau$$

$$= \frac{1}{\pi} \lim_{W\to\infty} \int_{-\infty}^{\infty} A_x\left(\frac{U}{W}\right) \frac{\sin U}{U}\, dU \quad \text{(letting } U = \tau W\text{)}$$

$$= \frac{1}{\pi} \int_{-\infty}^{\infty} A_x\left(\lim_{W\to\infty} \frac{U}{W}\right) \frac{\sin U}{U}\, dU$$

$$= \frac{1}{\pi} \int_{-\infty}^{\infty} A_x(0)\, \frac{\sin U}{U}\, dU = A_x(0)\, \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin U}{U}\, dU =$$

$$A_x(0) = \sigma_2.$$

Proof of L

In steady state, we can write, from A,

$$y(t) = \int_{0}^{\infty} W(U)\, x(t-U)\, dU \quad \text{and} \tag{2.2}$$

$$y(t+\tau) = \int_{0}^{\infty} W(V)\, x(t+\tau-V)\, dV.$$

Multiplying the above equations together,

$$y(t)\, y(t+\tau) = \int_{0}^{\infty} \int_{0}^{\infty} W(U)\, W(V)\, x\,(t-U)\, x\,(t+\tau-V)\, dU\, dV. \tag{2.3}$$

Taking expectations, t disappears from the right hand side (as will be seen from the following equalities) and we write:

(Integrate average rather than average integrals.)

$$(2.4)\qquad A_y(\tau) = E\int_0^\infty\int_0^\infty W(U)W(V)\ x\ (t-U)x\ (t+\tau-V)dU\ dV$$

$$= \int_0^\infty\int_0^\infty W(U)\ W(V)\ \ E[x(t-U)\ x\ (t+\tau-V)]\ dU\ dV$$

$$= \int_0^\infty\int_0^\infty W(U)\ W(V)\ A_x(\tau+U-V)\ dU\ dV.$$

y(t) is also stationary, and the fact that t disappeared is in agreement with this, although it is not a proof. The careful reader will note that we have not rigorously defined stationarity, and for obvious reasons we have therefore omitted the proof that y(t) is stationary.

Finally,

$$(2.5)\qquad \phi_y(S) = \int_{-\infty}^\infty e^{-S\tau}\ A_y(\tau)\ d\tau$$

$$= \int_{-\infty}^\infty e^{-S\tau}\left[\int_0^\infty\int_0^\infty W(U)\ W(V)\ A_x(\tau+U-V)\ dU\ dV\right] d\tau$$

$$= \int_0^\infty\int_0^\infty W(U)\ W(V)\ e^{-S(V-U)}\left[\int_0^\infty e^{-S(\tau+U-V)}\right.$$

$$\left. A_x(\tau+U-V)\ d\tau\right]\ dU\ dV$$

$$= \int_0^\infty\int_0^\infty W(U)\ W(V)\ e^{-SV}\ e^{SU}\ [\phi_x(S)]\ dU\ dV$$

$$= \phi_x(S)\left[\int_0^\infty W(U)\ e^{SU}\ dU\right]\left[\int_0^\infty W(V)\ e^{-SV}\ dV\right]$$

(Laplace transform of weighting function.)

$$= F(S)\ F(-S)\ \phi_x(S)\ .$$

The result for S = i follows from C.

Proof of M (Intuitive)

If $\phi_x(i\omega_o)$ were negative, we choose a stable transfer function F(S) such that $|F(i\omega)|^2$ is minutely small everywhere except near ω_o (and $-\omega_o$). Then $\frac{1}{2\pi}\int_{-\infty}^{\infty} |F(i\omega)|^2 \phi(i\omega)\, d\omega < 0$. If we send x(t) through the filter, the expected value of the square of the output, is, by K and L, just the above integral. On the other hand, the expected value of a square is never negative, a contradiction.

Discussion---Statements K and L are a key to the utility of the concept of P.S.D. If we know the P.S.D. of an input to a filter (which can be computed from the autocorrelation function by (2.1)) we know the P.S.D. and mean square of the output. Since missile motion can be considered as the output of a filter, one sees that we have a powerful tool for computing mean square miss distances when the P.S.D. or autocorrelation function of the input is known. Accuracy of tracking radars as well as the performance of other devices can be analyzed using these concepts.

In many practical cases, the integrals can be evaluated in closed form using residue theory, and tables exist from which they can be quickly computed.

Two examples of time series are target acceleration (say in one coordinate) and noise. Clearly target acceleration must generally be regarded as statistical, since the pilot is not going to tell us what he is going to do. Noise comes from many sources, and is always present to some degree in transmission and reception of radar energy. There is glint noise, fading noise, receiver noise, and others.

There are two types of P.S.D.'s which have been given names, and are worth mentioning. The first is $\phi(S) = \phi$, a constant, and the time series is said to be "white." One sometimes uses the phrase "white P.S.D." or "white noise." Such a time series cannot exist in nature, because σ^2 is infinite (from K). It is nonetheless a useful artifice. If $\phi(S)$ takes the form $\phi(S) = \frac{a^2}{b^2 - S^2}$, a, b real and different from zero, one substitutes the work "Markovian" for "white." Note that

$\phi(S)$ is the P.S.D. of the output of the filter $F(S) = \frac{|a|}{|b| + S}$ when the input has P.S.D. $= 1$. Although "white noise" cannot exist in nature, "Markovian noise" can.

The words "power spectral density" are well chosen. If a filter F(S) "filters out" all frequencies except those in a small interval, K and L show that the output power (σ_y^2) is proportional to $\phi(i\omega)$ for ω in the interval and proportional to the "bandwidth," or the length of the interval. Thus, P.S.D. is power per unit bandwidth. It can be seen that a white time series has the same power per unit bandwidth for all frequencies. A Markovian time series has a P.S.D. which looks like a white P.S.D. for small frequencies and decays away for large frequencies.

Exercise (d) If $\phi(S) = \frac{a^2}{b^2 - S^2}$, a, b real and different from zero, find σ^2.

III. OPTIMUM FILTERS AND THE WIENER HOPF EQUATION

Suppose the input to a filter is composed of the sum of two time series, x(t) + N(t), where x(t) is regarded as the "signal," or desirable part of the input, and N(t) is the "noise," or undesirable part of the input. Suppose that the P.S.D.'s or the autocorrelation functions of x(t) and N(t) are not known, and that x(t) and N(t) are statistically independent and have mean zero. This implies that

(3.1) $\quad E\,[x(t_1)\,N(t_2)] = 0$ for all t_1 and t_2.

We now wish to design the filter (with weighting function $W(\tau)$ and transfer function F(S)) which makes the output y(t) resemble the "signal" portion of the input, x(t), as closely as possible. The criterion used by Wiener is that the steady state value of $E\,(y(t) - x(t))^2$ should be a minimum. As we shall see, this leads us to a problem in calculus of variations. Let us denote the above expression by σ^2.

Then, since

(3.2) $$y(t) = \int_0^\infty W(\tau)\,[\,x(t - \tau) + N(t - \tau)\,]\,d\tau\,,$$

we have, by subtracting x(t) from both sides and squaring,

$$(3.3) \qquad [y(t) - x(t)]^2 = \int_0^\infty \int_0^\infty W(\tau)\, W(U)\, [x(t-\tau) + N(t-\tau)]\, [x(t-U) + N(t-U)]\, d\tau\, dU - 2 \int_0^\infty x(t)\, W(\tau)\, [\, x(t-\tau)\, d\tau + x(t)]^2 .$$

Let us denote the autocorrelation functions of x(t) and N(t) by $A_x(\tau)$ and $A_N(\tau)$ respectively, and the P.S.D.'s of x(t) and N(t) by $\phi_x(S)$ and $\phi_N(S)$, respectively. Averaging both sides of (3.3), we have

$$(3.4) \qquad \sigma^2 = \int_0^\infty \int_0^\infty W(\tau)\, W(U)\, [\, A_x(\tau - U) + A_N(\tau - U)\,]\, d\tau\, dU - 2 \int_0^\infty W(\tau)\, A_x(\tau)\, d\tau + A_x(0) .$$

<u>Exercise (e)</u> Prove that (Hint : Similar to Proof of K.)

$$\sigma^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left|1 - F(i\omega)\right|^2 \phi_x(i\omega)\, d\omega + \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| F(i\omega)\right|^2 \phi_N(i\omega)\, d\omega .$$

Our problem mathematically is to choose W(τ) or F(S) so that (3.4) or the expression in exercise (e) is minimum. The classical approach is to work with (3.4), although the

problem can be solved by working with the expression in exercise (e) directly. The latter approach is somewhat tricky, however, as care must be taken to optimize σ^2 over stable transfer functions. This is automatically taken care of in the first approach by virtue of the definition of weighting functions. We shall here use the first approach.

Suppose that $W(\tau)$ is such that (3.4) is a minimum. If $\eta(\tau)$ is an arbitrary fixed well behaved function of time, and ϵ an arbitrary number, we cannot then get a smaller value for σ^2 if we replace $W(\tau)$ by $W(\tau) + \epsilon\eta(\tau)$ in (3.4).

That is, if we write

$$(3.5)\qquad \sigma^2(\epsilon) = \int_0^\infty \int_0^\infty [W(\tau) + \epsilon\eta(\tau)][W(U) + \epsilon\eta(U)][A_x(\tau - U) + A_N(\tau - U)]\, d\tau\, dU$$

$$- 2\int_0^\infty [W(\tau) + \epsilon\eta(\tau)]\, A_x(\tau)\, d\tau + A_x(0).$$

Then $\sigma^2(\epsilon)$ must be a minimum for $\epsilon = 0$. So we differentiate (3.5) with respect to ϵ, and set $\epsilon = 0$, obtaining

$$(3.6)\qquad 0 = \int_0^\infty \int_0^\infty W(\tau)\,\eta(U)\,[A_x(\tau - U) + A_N(\tau - U)]\, d\tau\, dU$$

$$+ \int_0^\infty \int_0^\infty W(U)\,\eta(\tau)\,[A_x(\tau - U) + A_N(\tau - U)]\, d\tau\, dU$$

$$- 2\int_0^\infty \eta(\tau)\, A_x(\tau)\, d\tau\,.$$

Now the second integral is the same as the first, since $A_X(\tau - U) + A_N(\tau - U) = A_X(U - \tau) + A_N(U - \tau)$, so we put a 2 in front of the second integral and throw the first one away. Dividing by 2, we have

$$(3.7) \qquad \int_0^\infty \left\{ \int_0^\infty W(U) \left[A_X(\tau - U) + A_N(\tau - U) \right] dU - A_X(\tau) \right\} \eta(\tau)\, d\tau = 0 .$$

The expression within the braces must vanish identically for all positive values of τ, for if it did not vanish for $\tau = \tau_o$ say, one could choose a function $\eta(\tau)$ which was positive in the neighborhood of $\tau = \tau_o$ and zero elsewhere, and (3.7) would not then vanish. Since (3.7) must hold for arbitrary $\eta(\tau)$, we conclude that

$$(3.8) \qquad \int_0^\infty W(U) \left[A_X(\tau - U) + A_N(\tau - U) \right] dU = A_X(\tau) \text{ for } \tau > 0.$$

This is one form of the Wiener-Hopf Equation. Its significance is that the optimum weighting function must obey it.

Solving for $W(\tau)$ or $F(S)$ explicitly is somewhat tricky. We proceed as follows: Let $W(U)$ be 0 for negative U, and the lower limit 0 in the integral of (3.8) can then be replaced by $-\infty$.

Since (3.8) does not tell us what the integral is equal to for negative τ, we write

$$(3.9) \qquad \int_{-\infty}^\infty W(U) \left[A_X(\tau - U) \right] + A_N(\tau - U) \right] dU = A_X(\tau) + r(\tau)$$

where $r(\tau)$ vanishes for positive τ, and is otherwise unknown.

Next, we multiply (3.9) by $e^{-S\tau}$ and integrate from $-\infty$ to ∞.

If we set

(3.10) $R(S) = \int_{-\infty} e^{-S\tau}\, r(\tau)\, d\tau$, we have

(3.11) $$\phi_x(S) + R(S) = \int_{-\infty}^{\infty} e^{-S\tau} \left\{ \int_{-\infty}^{\infty} W(U)\, [A_x(\tau - U) + A_N(\tau - U)]\, dU \right\} d\tau$$

$$= \int_{-\infty}^{\infty} W(U)\, e^{-SU} \left\{ \int_{-\infty}^{\infty} e^{-S(\tau - U)} [A_x(\tau - U) + A_N(\tau - U)]\, d\tau \right\} dU$$

$$= \int_{-\infty}^{\infty} W(U)\, e^{-SU} \left\{ \int_{-\infty}^{\infty} e^{-SV}\, [A_x(V) + A_N(V)]\, dV \right\} dU$$

$$= \left\{ \int_{-\infty}^{\infty} W(U)\, e^{-SU}\, dU \right\} \left\{ \int_{-\infty}^{\infty} e^{-SV}\, [A_x(V) + A_N(V)]\, dV \right\}$$

$$= F(S)\ [\phi_x(S) + \phi_N(S)] \ .$$

So the optimum transfer function satisfies

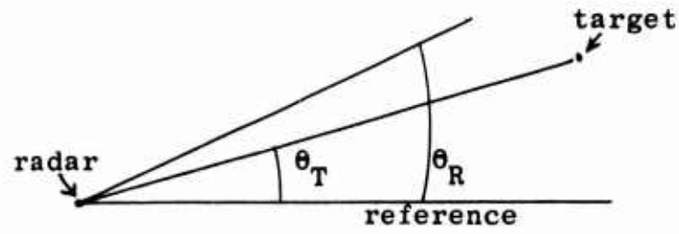
$$(3.12) \qquad F(S)\ [\phi_x(S) + \phi_N(S)] = \phi_x(S) + R(S)\ .$$

At first sight this looks like a useless expression since R(S) is unknown. It is not completely unknown however. Let us assume that $r(\tau)$ decays to zero as $\tau \to \infty$ sufficiently rapidly that the integral in (3.10) exists for S = i . From (3.9), this is reasonable if A_x and A_N are sufficiently well behaved. Inspection of (3.10) then shows that R(S) exists whenever the real part of S is negative. In practice, R(S) turns out to be a rational function. If it becomes infinite at all, it can only become infinite when the real part of S is positive. (Note that this is just the reverse for F(S).)

It turns out that this property of R(S) is sufficient to determine F(S) uniquely, using (3.12). The actual explicit expression for F(S) would take us far afield into notations for factors of $\phi_x(S) + \phi_N(S)$, removal of singular parts, and Liouville's theorem in analytic function theory. We shall not attempt to derive the general explicit form for F(S) here.

However, it will be seen in the next section how, in a given case, F(S) can be found.

IV. AN OPTIMIZATION EXAMPLE: DESIGN OF A TRACKING RADAR ANGLE LOOP



Suppose an aircraft is at an elevation angle Θ_T. When we attempt to measure Θ_T, however, we measure instead $\Theta_T + \Theta_N$, where Θ_N is noise. We now wish the radar direction,

Θ_R, to be as close to Θ_T as possible. Using the least squares criterion, the results of the last section can be seen to be applicable.

.. We shall assume that Θ_N has a white P.S.D. equal to ϕ, and Θ_T has a white P.S.D. equal to Θ. These are sometimes reasonable assumptions.

It is possible to get into trouble with these assumptions because Θ_T itself has a P.S.D. which is infinite for $S = 0$. Also $A_N(\tau)$ is infinite for $\tau = 0$. This means that some of the manipulations of the previous section which all assumed well-behaved functions, are not strictly legitimate.

We get around this by saying that ϕ_N has a P.S.D. $\frac{\phi}{1 - \epsilon^2 S^2}$ (instead of ϕ) and Θ_T has a P.S.D. $\frac{\Theta}{(S^2 - \delta^2)^2}$ (instead of $\frac{\Theta}{S^4}$) and eventually letting ϵ and $\delta \rightarrow$ zero (assume $\epsilon > 0$, $\delta > 0$) ..

From (3.12) then,

$$(4.1) \qquad F(S)\left[\frac{\Theta(1 - \epsilon^2 S^2) + \phi(S^2 - \delta^2)^2}{(1 - \epsilon^2 S^2)(S^2 - \delta^2)^2}\right] = \frac{\Theta}{(S^2 - \delta^2)^2} + R(S).$$

If we imagine both sides of (4.1) expressed as the sum of partial fractions, we see that R(S) must take the form

$$(4.2) \qquad R(S) = \frac{A}{S - \delta} + \frac{B}{(S - \delta)^2} + \frac{C}{1 - \epsilon S} + \text{polynomial}.$$

Let ϵ, $\delta \rightarrow 0$, we have

$$(4.3) \qquad F(S)\,\frac{\Theta + \phi S^4}{S^4} = \frac{\Theta}{S^4} + \frac{A}{S} + \frac{B}{S^2} + \text{polynomial},$$

or

$$(4.4) \qquad F(S) = \frac{\theta + BS^2 + AS^3 + S^4 \text{ (polynomial)}}{\theta + \phi S^4} \ .$$

From exercise (e) we see that the polynomial must be zero, for otherwise the second integral there would be infinite.

So, since the denominator factors,

$$(4.5) \qquad F(S) = \frac{\theta + BS^2 + AS^3}{(\sqrt{\theta} + \sqrt{2}\,\theta^{1/4}\phi^{1/4} S + \sqrt{\phi}\, S^2)\ (\sqrt{\theta} - \sqrt{2}\,\theta^{1/4}\,^{1/4} + \sqrt{\phi}\ S^2)} \ .$$

Since F(S) can become infinite only for values of S which have negative real parts, the second factor of the denominator must be an exact divisor of the numerator. If one attempts this division, one will discover that this can only happen if

$$(4.6) \qquad \theta + BS^2 + AS^3 = (\sqrt{\theta} + \sqrt{2}\,\theta^{1/4}\phi^{1/4}S)\ (\sqrt{\theta} - \sqrt{2}\,\theta^{1/4}\phi^{1/4}S + \sqrt{\phi}\ S^2)$$

Hence

$$(4.7) \qquad F(S) = \frac{\sqrt{\theta} + \sqrt{2}\,\theta^{1/4}\phi^{1/4}S}{\theta + \sqrt{2}\,\theta^{1/4}\phi^{1/4} S + \sqrt{\phi}S^2}$$

This is commonly written in the form

$$(4.8) \qquad F(S) = \frac{1 + 2\zeta\tau S}{1 + 2\zeta\tau S + \tau^2 S^2}$$

where

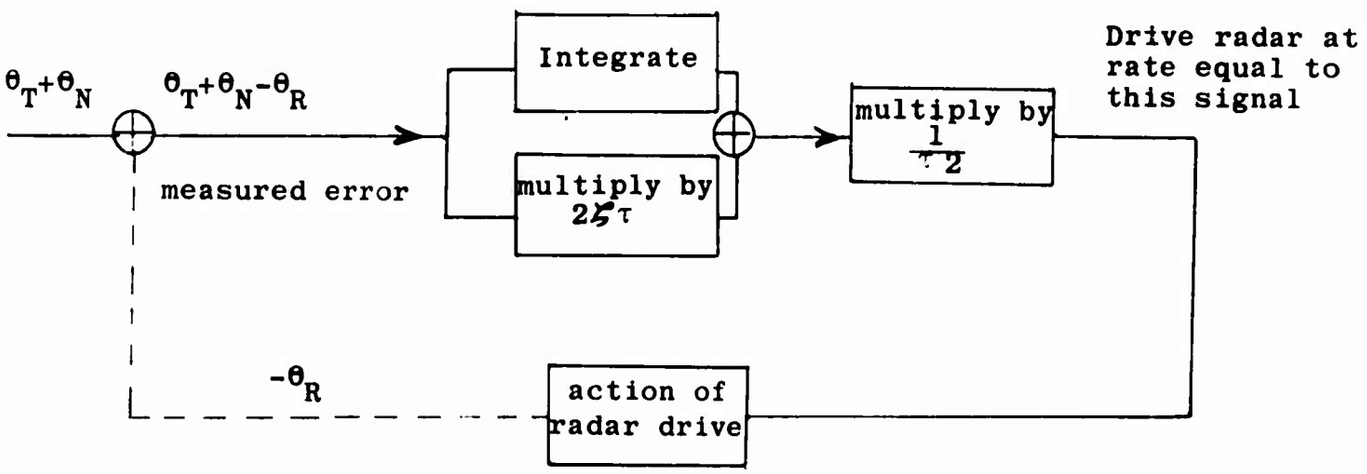
$$(4.9) \qquad \zeta = \frac{1}{\sqrt{2}} \ , \quad \tau = \frac{\phi^{1/4}}{\theta^{1/4}} \ .$$

From exercise (e) and the use of tables, the minimum value of σ^2 can be computed to be

$$(4.10) \qquad \sigma^2 = \sqrt{2}\ \phi^{3/4}\theta^{1/4} \ .$$

The problem is not completely finished, however, as the radar does not actually measure $\theta_T + \theta_N$, but $\theta_T + \theta_N - \theta_R$. By the action of the electromagnetic propagation, the measurement is in the form of boresight angular error. This measured error must then be used to physically drive the radar as a servomechanism until the error is nulled.

The operation of the radar might then be as indicated in the following diagram



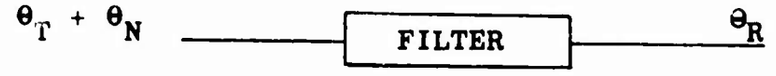
Hence

$$(4.11) \qquad \dot{\theta}_R = \frac{1}{\tau^2}\left[2\tau\,(\theta_T + \theta_N - \theta_R) + \int (\theta_T + \theta_N - \theta_R)\,dt\right]$$

or

$$(4.12) \qquad \tau^2\,\ddot{\theta}_R + 2\tau\,\dot{\theta}_R + \theta_R = 2\tau\,(\dot{\theta}_T + \dot{\theta}_N) + (\theta_T + \theta_N) \ .$$

Hence, we have the equivalent diagram:



where the transfer function of the filter is given by (4.7) or (4.8). A tracking radar behaving as just described will, in the least squares sense, then track the target as closely as possible, whenever the statistical assumptions are justified. Other statistical assumptions will, of course, lead to different answers.

Exercise (f) If x(t) is a stationary time series, show that

$$A_x(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega\tau} \phi_x(i\omega)\, d\omega.$$ (Hint: see Proof of K.)

Exercise (g) In the example of section IV, assume instead that θ_T is white. Find the optimum transfer function, the optimum mean square tracking error, and draw the radar block diagram.

Exercise (h) What is the weighting function of a filter whose transfer function is $\frac{1}{1+TS}$?

Exercise (i) Let x(t) be a time series such that $A_x(\tau) = ke^{-\beta|\tau|}$. Show that x(t) is Markovian. Use this result and exercise (f) to evaluate the integral $\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\cos \omega\tau}{1 + T^2 \omega^2}\, d\omega$.

SUMMARY AND DISCUSSION

We have here attempted to introduce the concepts of filter, transfer function, weighting function, time series, stationarity, ergodicity, autocorrelation function, power spectral density and noise. We have also introduced the Wiener theory of optimization, derived the Wiener-Hopf equation, and have shown how it can be solved. A particular application has been discussed in detail, and a tracking radar angle loop has been optimized in the presence of an input which has been corrupted with noise.

The Wiener theory can be applied to other cases, although the derivations in this paper must be somewhat altered to achieve greater generality. An example is the design of an optimum loop when some parts of the loop have already specified transfer functions. This could occur in the design of a beam riding guidance computer where the aerodynamic equations have already been determined by the airframe design. Wiener theory can also be applied (by the use of Lagrange multipliers) to problems with constraints. A typical problem of this type is the design of a beam riding guidance computer to minimize deviation from the line of sight subject to the constraint that the mean square acceleration is not too large. (Or else the missile might slow down too much or fall apart.)

In general, Wiener theory is applicable only to steady state problems where the time series may be considered as stationary. Because of the mathematical simplifications which occur for this case, the theory is quite well developed. Although a considerable amount of work has been done on practical cases for which Wiener theory does not apply, the general case is nowhere near as well understood. For example, transient statistical problems, non-stationary problems, final value problems (in which one is interested in the output only at a specified time) all occur in missile work, and seem to require special handling.

Wiener theory is directly applicable to a large fraction of statistical design problems in missile work, nonetheless. Some people prefer to take a different approach to some of these problems. One such approach is the game theory approach, where it is assumed that the target maneuvers in such a way as to maximize the tracking error (or miss distance), and the filter is designed to minimize this maximum.

Another different class of problems occurs when the time series under consideration are defined discretely, i.e., say at multiples of Δt. This type of problem occurs with search radars, which only "look" at a target every so often. One is then concerned with maintaining track on this target and minimizing the tracking error. A system which tracks in this manner is referred to by some as a "track while scan" system.

Finally, it should be stated that although the results of section I are prerequisites to the subsequent sections, they are extremely important in their own right. Because the emphasis of this paper has been on time series and optimization in the presence of noise, it might appear that section I was

just the means to an end. Such is not the case. The concept and properties of a transfer function are to most missile engineers as the ABC's are to a school child. They use them almost every working day in a practical sense, even though some of them may know nothing of time series or Wiener theory.

It is safe to say that the science of guided missiles would be far behind its present state of development if it were not for the concept of a transfer function. It is therefore strongly suggested that the reader absorb section I thoroughly, whether or not he has the time or patience to absorb the rest of the paper. Section II is of secondary importance, sections III and IV of less importance still.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

PERTURBATION METHODS

by

S. T. Haywood

I. A WORD ON TAYLOR'S SERIES

If $y = f(x)$ is a twice differentiable function defined on an interval $a \leq x \leq b$ containing a point x_o, then

$$(1) \qquad y = f(x) = f(x_o) + f'(x_o)\,(x-x_o) + \frac{1}{2} f''(\xi)\,(x-x_o)^2 ,$$

where ξ is a point between x and x_o. This is a form of Taylor's series with remainder. The remainder term, namely

$$(2) \qquad \frac{1}{2} f''(\xi)\,(x-x_o)^2 ,$$

represents the error committed when the first two terms alone are used as an approximation to $f(x)$.

$$(3) \qquad \text{Error} = \left| f(x) - f(x_o) - f'(x_o)\,(x-x_o) \right| = \frac{1}{2} \left| f''(\xi) \right| (x-x_o)^2 .$$

The number ξ usually is itself a complicated function of x. Normally, the only thing which can be said with assurance about ξ is that it lies <u>between</u> x and x_o, and hence lies in the interval $[a, b]$. If we let M be the maximum of $|f''(x)|$ for $a \leq x \leq b$, then $|f''(\xi)| \leq M$ and we may write

$$(4) \qquad \text{Error} = \left| f(x) - f(x_o) - f'(x_o)\,(x-x_o) \right| \leq \frac{M}{2} (x-x_o)^2 .$$

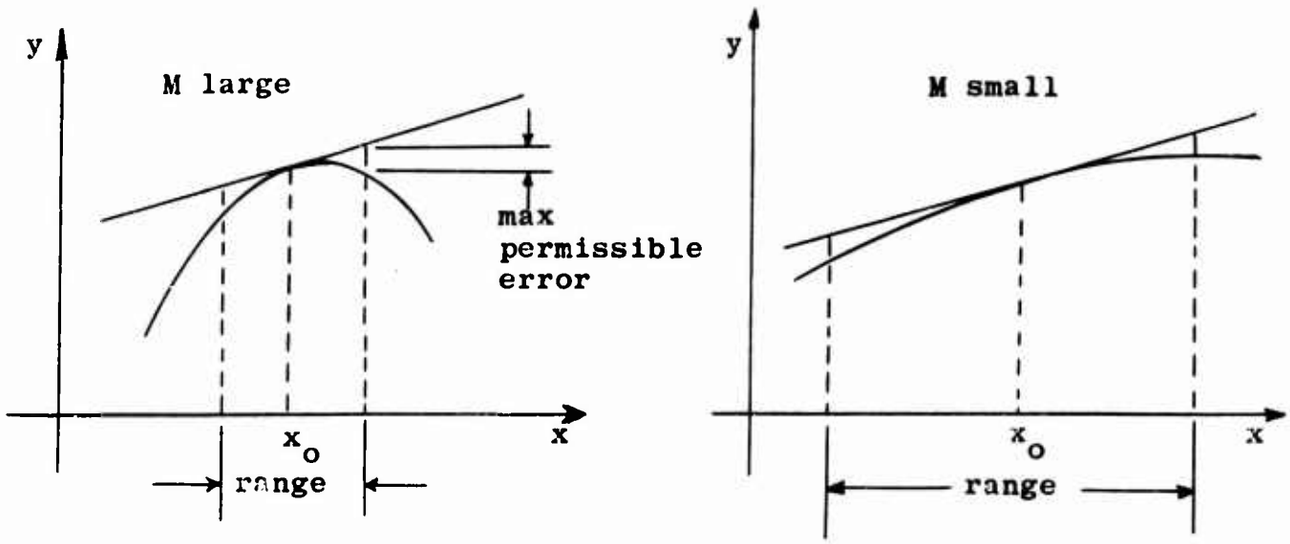
Thus, although it is usually very difficult to find the <u>exact</u> error, it is fairly easy to establish a <u>bound</u> for the error. Usually this is quite sufficient.

The form of Eq. (4) shows that the error will be small if x is "close enough" to x_o. Just how close is "close enough" depends on how much error is permissible and on the magnitude of M.

II. LINEAR APPROXIMATIONS

$$f(x) \cong f(x_o) + f'(x_o)\,(x-x_o) \tag{5}$$

is called the "linear approximation to $f(x)$ at $x = x_o$" inasmuch as the right hand side is linear. Geometrically, the approximating function is the tangent to $y = f(x)$ at $x = x_o$. From the preceding discussion, it is evident that there is a range of values of x "close" to x_o for which the linear approximation is adequate, i.e., the error committed by using the approximation lies within acceptable bounds. The length of the range depends on the behavior of $f(x)$ in the neighborhood of $x = x_o$. More specifically, if $f''(x)$ is large near $x = x_o$ (which means that the curvature is large), then M will be large and the length of the range will small (for a given permissible error). On the other hand if $f''(x)$ is small near $x = x_o$ (small curvature), then the range will be correspondingly large.



For functions of more than one variable the results are quite similar. For example, the linear approximation to $w = w(x, y, z)$ at $(x, y, z) = (x_o, y_o, z_o)$ is

$$(6) \quad w(x,y,z) \cong w(x_o,y_o,z_o) + A(x-x_o) + B(y-y_o) + C(z-z_o)$$

where

$$(7) \quad A = \frac{\partial w}{\partial x}$$

$$(8) \quad B = \frac{\partial w}{\partial x}$$

$$(9) \quad C = \frac{\partial w}{\partial x}$$

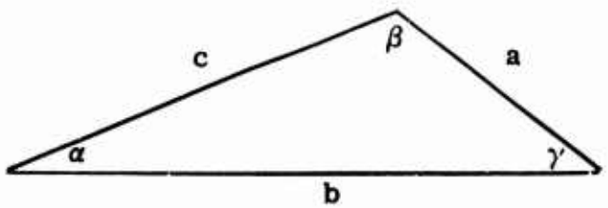
evaluated at $x=x_o$, $y=y_o$, $z=z_o$.

As before, there is a set of points (x, y, z) around (x_o, y_o, z_o) for which the error in the approximation is small enough to be acceptable. Once again, the actual extent of the point set depends on the values of the second partial derivatives in the vicinity of (x_o, y_o, z_o).

In the case of a function of two variables, the geometrical interpretation of the linear approximation is readily available. The approximating function is merely the tangent plane to the surface $w = f(x,y)$ at (x_o, y_o).

III. FORMULATION OF PROBLEMS

The mathematical form of an engineering problem is an equation or a set of equations whose solution gives the desired answers. Consider, for example, the following simple problem which could arise in connection with surveying. Find the angles of a triangle whose three sides are known.



The mathematical form of this particular problem can be written as the following three equations.

(10) $a^2 = b^2 + c^2 - 2bc \cos \alpha$ (law of cosines),

(11) $\frac{\sin \beta}{b} = \frac{\sin \alpha}{a}$ (law of sines),

(12) $\alpha + \beta + \gamma = \pi$.

It should be noted that the mathematical form of the problem is not unique. Equation (11), for instance, could be replaced by

(13) $b^2 = a^2 + c^2 - 2ac \cos \beta$ (law of cosines).

IV. EQUATIONS CONTAINING PARAMETERS

The solution of an equation or a set of equations usually will depend on one or more parameters. In the surveying problem just considered, the solution depends on three parameters, namely the sides of the traingle a, b, c. Other examples of this are given in the following equations.

(14) $x^3 - ax + 6 = 0$ — 1 parameter --- a ,

(15) $\begin{cases} ax^2 + y^2 = 25 \\ x + by = 5 \end{cases}$ — 2 parameters -- a, b,

(16) $e^{1-x} = ax$ — 1 parameter --- a ,

(17) $\begin{cases} y + (3 + ay^2)\,\dot{y} + 2y = 0 \\ y = 1,\ \dot{y} = -1 \text{ when } t = 0 \end{cases}$ — 1 parameter --- a ,

(18) $\begin{cases} x\dot{y} = 1 \\ y\dot{x} = ax - 1 \\ t = o,\ x = 1,\ y = 1 \end{cases}$ — 1 parameter --- a .

The first three examples (Eqs. 14, 15, and 16) are ordinary equations for which the solutions are numbers. The last two examples (Eqs. 17 and 18) are differential equations for which the solutions are functions satisfying certain initial conditions.

V. METHOD OF LINEAR PERTURBATIONS

The following situation very frequently occurs. The solution to an equation is known for certain values of the parameters appearing in the equation. It is desired to find the solution when the parameters have values differing slightly from the values for which the solution is known. There is a fairly standard terminology to describe this state of affairs. "Knowing the solution for given values of the parameters, find the behavior of the solution when the parameters are perturbed." The word perturbed carries the

connotation of differing slightly.

One method for handling the problem just described is the method of linear perturbations, for which the groundwork has been laid in the preceding paragraphs. The idea is quite simple. The solution to the equation is a function of the parameters appearing in the equation. The value of this function is known for certain given values of the parameters. The method of linear perturbations consists of replacing the function by its linear approximation at the known point. Consider the following example, which illustrates the method.

VI. EXAMPLES

Let the equation to be solved be

$$x^2 + a = 8x . \tag{19}$$

This equation has one parameter, a. When a = 12, the equation has the solutions x = 2 and x = 6. What will be the solutions for values of a close to a = 12? The solutions are functions of a. Indeed, since the equation is a quadratic, we can easily find

$$x = 4 \pm \sqrt{16 - a}. \tag{20}$$

We shall consider only one of these solutions, namely

$$x = x(a) = 4 + \sqrt{16 - a}. \tag{21}$$

For this case we have

$$x(12) = 6. \tag{22}$$

Furthermore,

$$\frac{dx}{da} = -\frac{1}{2\sqrt{16 - a}} \quad , \text{ so that} \tag{23}$$

$$\left(\frac{dx}{da}\right)_{a=12} = -\frac{1}{4} \, . \tag{24}$$

The linear approximation to $x(a)$ at $a = 12$ is

$$x(a) \cong x(12) + \left(\frac{dx}{da}\right)_{a=12} (a-12) = 6 - \frac{a-12}{4} = 9 - \frac{a}{4} \, . \tag{25}$$

Equation (25) now can be used to estimate the solution for values of a near $a = 12$. Thus, when $a = 12.04$, we obtain $x \cong 5.99$. The exact result, of course, can be obtained from Eq. (21) which gives $x = 4 + \sqrt{3.96} \cong 5.989975$ (making use of a table of square roots). We now see that the error in the linear approximation is about $5.99 - 5.989975 = 0.000025$.

The preceding example is so simple that it is apt to give rise to certain misconceptions which it would be well to dispel immediately. Simple as Eq. (21) is, Eq. (25) is even more simple, so that there is a clear advantage in using Eq. (25) rather Eq. (21). Nevertheless, Eq. (21) is simple enough so that there can be no real objection to its use. Besides, apparently we need Eq. (21) anyway in order to get the derivative which we used in Eq. (25). As a matter of fact, this last statement is false. We do not need an explicit representative of $x(a)$, as in Eq. (21), in order to find the derivative

$\frac{dx}{da}$. This is very fortunate, since usually it is impossible to find such a representation. In such a case we use <u>implicit</u> differentiation to find the required derivatives.

This technique would be applied to our problem in the following manner. Equation (19), namely

$$x^2 + a = 8x \tag{19}$$

is differentiated implicitly with respect to <u>a</u>, observing that $x = x(a)$, to give

$$2x \frac{dx}{da} + 1 = 8 \frac{dx}{da} , \quad \text{whence} \tag{26}$$

$$\frac{dx}{da} = - \frac{1}{2(x-4)} . \tag{27}$$

Setting $a = 12$ and $x = 6$ (remember that it is <u>known</u> that $x = 6$ when $a = 12$) gives

$$\left.\frac{dx}{da}\right|_{a=12} = - \frac{1}{4} . \tag{24}$$

Observe that Eq. (21) did not appear at all in this method.

Next, we consider a case involving two parameters. The equations are

$$ax^2 + y^2 = 25 \tag{15}$$

$$x + by = 5$$

and it is desired to find the behavior of the solutions in the vicinity of $a = 1$, $b = 3$. As in our previous example, it is possible to express both x and y explicitly as functions of a and b, since nothing worse than a quadratic equation is involved; however, we shall avoid this approach. First of all, the solutions when $a = 1$, $b = 3$ are $x = -4$, $y = 3$ and $x = 5$, $y = 0$. We shall treat the case $x = -4$, $y = 3$.

$$(28)\quad \begin{cases} x = x(a,b) \cong x(1,3) + A(a-1) + B(b-3) \\ y = y(a,b) \cong y(1,3) + C(a-1) + D(b-3) \end{cases} \qquad \text{where}$$

$$(29)\quad A = \frac{\partial x}{\partial a},\ B = \frac{\partial x}{\partial b},\ C = \frac{\partial y}{\partial a},\ D = \frac{\partial y}{\partial b}$$

and the derivatives are to be evaluated at $a = 1$, $b = 3$. The problem now consists of computing A, B, C, D. Differentiating Eq. (15) partially with respect to $\underline{a}$ and $\underline{b}$ yields

$$(30)\quad \begin{cases} 2ax\dfrac{\partial x}{\partial a} + x^2 + 2y\dfrac{\partial y}{\partial a} = 0 \\ \dfrac{\partial x}{\partial a} + b\dfrac{\partial y}{\partial a} = 0 \\ 2ax\dfrac{\partial x}{\partial b} + 2y\dfrac{\partial y}{\partial b} = 0 \\ \dfrac{\partial x}{\partial b} + b\dfrac{\partial y}{\partial b} + y = 0\ . \end{cases}$$

When these are evaluated at $a = 1$, $b = 3$, $x = -4$, $y = 3$, the result is

$$(31)\quad \begin{cases} -8A + 16 + 6C = 0 \\ A + 3C = 0 \\ -8B + 6D = 0 \\ B + 3D + 3 = 0. \end{cases}$$

These equations are easily solved to give

$$(32) \qquad A = \frac{8}{5}\ ,\ B = -\frac{3}{5}\ ,\ C = -\frac{8}{15}\ ,\ D = -\frac{4}{5}\ .$$

Hence

$$(33) \qquad \begin{cases} x \cong -4 + \frac{8}{5}(a-1) - \frac{3}{5}(b-3) \\ y \cong 3 - \frac{8}{15}(a-1) - \frac{4}{5}(b-3) \end{cases}$$

VII. PERTURBATIONS APPLIED TO DIFFERENTIAL EQUATIONS

One of the main applications of perturbation theory is to systems of equations consisting partly or wholly of differential equations. As an example, we consider

$$(18) \qquad \begin{aligned} x\dot{y} &= 1 \\ y\dot{x} &= ax - 1 \end{aligned} \qquad x = y = 1 \qquad \text{when } t = 0\ .$$

We assume the parameter a is small in absolute value. When $a = 0$, the solution to the above equations is

$$(34) \qquad \begin{cases} x = e^{-t} \\ y = e^{t}\ . \end{cases}$$

How should this be modified when $a = 0$? We acknowledge that x and y, as well as being functions of t, will also depend on the parameter a. Hence, we write $x = x(t,a)$, $y = y(t,a)$. The linear approximations to these functions in the neighborhood of $a = 0$ are

$$(35)\qquad \begin{cases} x(t, a) \cong x(t, o) + Aa \\ y(t, a) \cong y(t, o) + Ba \end{cases} \qquad \text{where}$$

$$(36)\qquad \begin{cases} x(t, o) = e^{-t} \\ y(t, o) = e^{t} \end{cases} \qquad \text{(see Eq. 34) and}$$

$$(37)\qquad \begin{cases} A = \dfrac{\partial x}{\partial a} \\ B = \dfrac{\partial y}{\partial a} \end{cases} \qquad \text{evaluated at } a = 0 \ .$$

It should be observed that A and B ordinarily will be functions of t. Our problem now is to compute these derivatives.

The technique we use is to observe that x(t, a), y(t, a) must satisfy Eq. (18) identically, that is

$$(38)\qquad \begin{cases} x \dfrac{\partial y}{\partial t} = 1 \\ y \dfrac{\partial x}{\partial t} = ax - 1 \ . \end{cases}$$

Differentiating these equations partially with respect to $\underline{a}$ yields

$$(39)\qquad \begin{cases} x \dfrac{\partial^2 y}{\partial a \, \partial t} + \dfrac{\partial x}{\partial a} \dfrac{\partial y}{\partial t} = 0 \\ y \dfrac{\partial^2 x}{\partial a \, \partial t} + \dfrac{\partial y}{\partial a} \dfrac{\partial x}{\partial t} = a \dfrac{\partial x}{\partial a} + x \ . \end{cases}$$

We now set $a = 0$, $x = e^{-t}$, $y = e^{t}$, $\frac{dx}{dt} = -e^{-t}$, and $\frac{dy}{dt} = e^{t}$ to obtain

$$(40) \qquad \begin{cases} e^{-t}\frac{dB}{dt} + e^{t}A = 0 \\ e^{t}\frac{dA}{dt} - e^{-t}B = e^{-t} \end{cases}$$

which are the equations we must solve in order to find A and B. It is evident that our procedure so far is the same as it was in our previous examples. Here we must contend with differential equations, and the equations we must finally solve, i.e., Eq. (40), are themselves differential equations, whereas our final equations were algebraic in the other examples. Nevertheless, the technique for finding those equations is the same in both cases.

Since Eq. (40) is a differential equation, it may seem that no great advantage has been derived by use of the perturbation analysis, but this would indeed be a false evaluation. Equation (40) is linear, whereas the original Eq. (18) is nonlinear. This conversion from nonlinear to linear (quite naturally called linearization) is the direct result of using the linear approximations. It is a decided advantage for the home team to have linear equations to solve.

VIII. INITIAL CONDITIONS

In order to solve Eq. (40), we require initial conditions for A and B. These are obtained by using the initial conditions appearing with the original Eq. (18). There we see that $x = y = 1$ when $t = 0$, no matter what the value of a is. Hence, setting $t = 0$ in Eq. (35) gives $A = 0$, $B = 0$. These are the initial conditions to be used in conjunction with Eq. (40). The solutions are

$$(41) \qquad \begin{cases} A = t\,e^{-t} \\ B = (1 - t)\,e^{t} - 1\,, \quad \text{whence} \end{cases}$$

$$(42) \qquad \begin{cases} x = e^{-t} + at\, e^{-t} \\ y = e^{t} + a\,[(1 - t)e^{t} - 1] \end{cases} .$$

These, then, are the answers we set out to obtain.

Now it happens that Eq. (18) can be solved explicitly for x and y in terms of t and a. The results are

$$(43) \qquad \begin{cases} x = e^{-(1-a)t} \\ y = \dfrac{e^{(1-a)t} - a}{1 - a} \end{cases}$$

which, for a = 0, reduce to Eq. (36). It should be kept in mind, however, that it is very seldom possible to get such explicit analytic solutions. We introduce Eq. (43) merely to illustrate another point. First, we consider two infinite series.

$$(44) \qquad e^{z} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots \quad \text{valid for all } z,$$

$$(45) \qquad \frac{1}{1-z} = 1 + z + z^2 + z^3 + \ldots \quad \text{valid for } |z| < 1$$

Thus, by applying Eq. (44) to Eq. (43), we obtain

$$(46) \qquad x = e^{-(1-a)t} = e^{-t}\, e^{at} = e^{-t}\left[1 + at + \frac{(at)^2}{2!} + \ldots\right] .$$

Similarly, we obtain

$$(47) \qquad y = [1 + a + a^2 + \ldots]\left\{e^{t}\left[1 - at + \frac{(at)^2}{2!} - \ldots\right] - a\right\} .$$

First consider Eq. (46). The series $1 + \frac{(at)^2}{2!} + \ldots$ converges for all values of a and t. In addition, whenever the product at is small enough, the first two terms of the series, i.e., the linear approximation $1 + at$, can be used without excessive error. Hence, Eq. (46) yields

$$x \cong e^{-t}(1 + at) = e^{-t} + at\,e^{-t} \tag{48}$$

if at is sufficiently small. This is precisely the result obtained by the perturbation analysis. (Compare with Eq. (42).) The present approach should make it clear that, not only must a be close to zero, but t must be sufficiently small also in order for the perturbation analysis to remain valid. In short, the point we wish to make is that, when applying perturbation methods to differential equations, it must be kept in mind that the results are apt to be valid only for a restricted range of the independent variable, not to mention the restricted ranges of the parameters. This does not prevent the method from being a very useful one.

Equation (47) also can be simplified under the assumption that at and a are small.

$$y \cong (1 + a)\,[e^t(1 - at) - a] = (1 + a)\,[e^t - a(te^t + 1)] \tag{49}$$

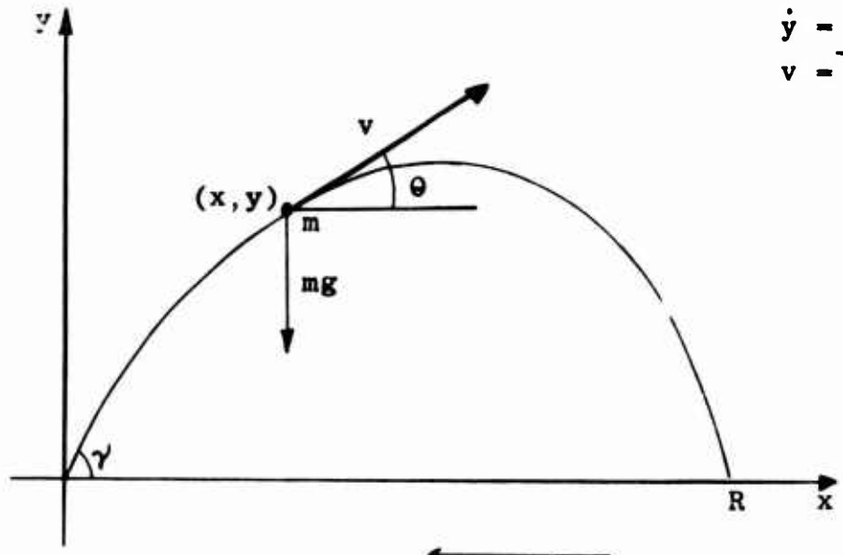
$$\cong e^t + a\,[e^t - (te^t + 1)] = e^t + a\,[(1 - t)\,e^t - 1] .$$

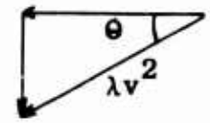
Once again this is the result given previously in Eq. (42). Clearly the procedure employed is to discard all powers higher than the first power of the small quantity. This automatically leads one to the linear approximation.

IX. A TRAJECTORY PROBLEM

As another example, we consider a simplified version of a fairly complicated problem. A missile of mass m is fired at an angle of elevation γ with a muzzle velocity b. Thereafter, it travels in free flight subject only to gravity and to air resistance which is proportional to the square of the velocity.

$$\dot{x} = v \cos\theta$$
$$\dot{y} = v \sin\theta$$
$$v = \sqrt{\dot{x}^2 + \dot{y}^2}$$





Air resistance force diagram

The equations of motion (obtained from Newton's Law) are

$$(50)\quad \begin{cases} \ddot{x} = -a\dot{x}\sqrt{\dot{x}^2 + y^2} \\ \ddot{y} = -a\dot{y}\sqrt{\dot{x}^2 + \dot{y}^2} - g \end{cases} \qquad t = 0 \qquad \begin{cases} x = y = 0 \\ \dot{x} = b\cos\gamma, \\ \dot{y} = b\sin\gamma. \end{cases}$$

$$a = \frac{\lambda}{m}$$

One of our simplifications will be to assume that the earth is flat over the distance travelled by the missile. The main simplification will be to assume that a is small, i.e., $a = 0$. The solution of Eq. (50) depends on three parameters, a, b, γ, and may be written as

$$\begin{cases} x = x\,(t, a, b, \gamma) \\ y = y\,(t, a, b, \gamma) \end{cases} . \tag{51}$$

The time of flight, T, of the missile is the time required for y, which started at zero, to reach zero once again. Hence T is defined by

$$y(T, a, b, \gamma) = 0 . \tag{52}$$

When solved for T, this yields

$$T = T\,(a, b, \gamma) . \tag{53}$$

The range, R, of the missile is the value of x when $t = T$. Hence R is defined by

$$R = x\,(T, a, b, \gamma) \tag{54}$$

whence, in view of Eq. (53),

$$R = R\,(a, b, \gamma) . \tag{55}$$

Thus, T and R are functions of the three parameters, a, b, γ. When $a = 0$, $b = b_o$, $\gamma = \gamma_o$, it is very easy to solve all the equations involved and find the values of T and R. We want to find out how T and R behave when the parameters are perturbed.

The linear approximations to Eqs. (53) and (55) are

$$(56)\quad \begin{cases} T \cong T(0, b_o, \gamma_o) + Aa + B(b - b_o) + C(\gamma - \gamma_o) \\ R \cong R(0, b_o, \gamma_o) + Ea + F(b - b_o) + G(\gamma - \gamma_o) \end{cases} \quad \text{where}$$

$$(57)\quad \begin{cases} A = \dfrac{\partial T}{\partial a} & E = \dfrac{\partial R}{\partial a} \\ B = \dfrac{\partial T}{\partial b} & F = \dfrac{\partial R}{\partial b} \\ C = \dfrac{\partial T}{\partial \gamma} & G = \dfrac{\partial R}{\partial \gamma} \end{cases} \quad \text{evaluated at } a = 0,\ b = b_o,\ \gamma = \gamma_o \ .$$

We must compute the derivatives A, B, C, E, F, G. For this we must turn to Eqs.(53) and (54) which define T and R implicitly. Differentiating Eqs. (52) and (54) partially with respect to a, b, γ, we obtain

$$(58)\quad \begin{cases} \dfrac{\partial y}{\partial T}\dfrac{\partial T}{\partial a} + \dfrac{\partial y}{\partial a} = 0 & \dfrac{\partial x}{\partial T}\dfrac{\partial T}{\partial a} + \dfrac{\partial x}{\partial a} = \dfrac{\partial R}{\partial a} \\ \dfrac{\partial y}{\partial T}\dfrac{\partial T}{\partial b} + \dfrac{\partial y}{\partial b} = 0 & \dfrac{\partial x}{\partial T}\dfrac{\partial T}{\partial b} + \dfrac{\partial x}{\partial b} = \dfrac{\partial R}{\partial b} \\ \dfrac{\partial y}{\partial T}\dfrac{\partial T}{\partial \gamma} + \dfrac{\partial y}{\partial \gamma} = 0 & \dfrac{\partial x}{\partial T}\dfrac{\partial T}{\partial \gamma} + \dfrac{\partial x}{\partial \gamma} = \dfrac{\partial R}{\partial \gamma} \end{cases} .$$

X. THE KNOWN SOLUTION

We digress momentarily to consider the case $a = 0$, $b = b_o$, $\gamma = \gamma_o$, and to compute some quantities which we shall presently need. We let

$$(59) \qquad \begin{cases} X = x(t, 0, b_o, \gamma_o) \\ Y = y(t, 0, b_o, \gamma_o) \end{cases}$$

and observe that these functions satisfy

$$(60) \qquad \begin{cases} \dot{X} = 0 \\ \ddot{Y} = -g \end{cases} \qquad t = 0 \quad \begin{cases} X = Y = 0 \\ \dot{X} = b_o \cos \gamma_o, \ \dot{Y} = b_o \sin \gamma_o. \end{cases}$$

These equations are particularly easy to solve. The results are

$$(61) \qquad \begin{cases} X = b_o\, t \cos \gamma_o & \dot{X} = b_o \cos \gamma_o \\ Y = b_o\, t \sin \gamma_o - \frac{1}{2} g t^2 & \dot{Y} = b_o \sin \gamma_o - gt. \end{cases}$$

From these we easily obtain

$$(62) \qquad \begin{cases} T_o = T(0, b_o, \gamma_o) = \dfrac{2 b_o \sin \gamma_o}{g} \\ R_o = R(0, b_o, \gamma_o) = \dfrac{b_o^2 \sin 2\gamma_o}{g} \end{cases} .$$

Hence, when $t = T_o$, we find

$$(63) \qquad \begin{cases} \dot{X} = b_o \cos \gamma_o \\ \dot{Y} = - b_o \sin \gamma_o \end{cases} .$$

Return now to Eq. (58). These must be evaluated at $a = 0$, $b = b_o$, $\gamma = \gamma_o$, and therefore at $T = T_o$ also. Under these conditions, the derivatives $\partial y/\partial T$ and $\partial x/\partial T$ become the expressions given in Eq. (63). However, there still remain the derivatives $\partial y/\partial a$, $\partial y/\partial b$, etc., about which, as yet, we have no information. Hence, we let

$$(64) \qquad \begin{cases} H = \dfrac{\partial x}{\partial a} = H(t) & P = \dfrac{\partial y}{\partial a} = P(t) \\ M = \dfrac{\partial x}{\partial b} = M(t) & Q = \dfrac{\partial y}{\partial b} = Q(t) \\ N = \dfrac{\partial x}{\partial \gamma} = N(t) & W = \dfrac{\partial y}{\partial \gamma} = W(t) \end{cases} \qquad \begin{array}{c} \text{evaluated at} \\ a = 0 \\ b = b_o \\ \gamma = \gamma_o \end{array} .$$

The evaluation of Eq. (58) now gives

$$(65) \qquad \begin{cases} - b_o \sin \gamma_o A + P(T_o) = 0 & b_o \cos \gamma_o A + H(T_o) = E \\ - b_o \sin \gamma_o B + Q(T_o) = 0 & b_o \cos \gamma_o B + M(T_o) = F \\ - b_o \sin \gamma_o C + W(T_o) = 0 & b_o \cos \gamma_o C + N(T_o) = G \end{cases} .$$

It is a very simple matter now to solve this set of simultaneous linear equations for A, B, C, E, F, G. Unfortunately, our troubles have merely been transferred to another place. We now find it necessary to evaluate $P(T_o)$, $Q(T_o)$, etc. These quantities are partial derivatives of the trajectory variables x and

y with respect to the parameters a, b, γ, and presumably can be obtained by differentiation of Eq. (51). However, we do not actually have the explicit representation, Eq. (51). All we have is the original set of differential equations (50) which define x and y. By differentiating Eq. (50) partially with respect to a, b, γ, and then evaluating at a = 0, $b = b_o$, $\gamma = \gamma_o$, we obtain

$$(66) \quad \begin{cases} \dfrac{d^2H}{dt^2} = -\dot{X}V & \dfrac{d^2P}{dt^2} = -\dot{Y}V \\ \dfrac{d^2M}{dt^2} = 0 & \dfrac{d^2Q}{dt^2} = 0 \\ \dfrac{d^2N}{dt^2} = 0 & \dfrac{d^2W}{dt^2} = 0 \ , \end{cases} \qquad \text{where}$$

$$(67) \quad V = \sqrt{\dot{X}^2 + \dot{Y}^2} \ .$$

This is a set of linear differential equations, in contrast to Eq. (50) which is nonlinear. We now require initial conditions for Eq. (66) in order to solve them.

XI. INITIAL CONDITIONS

The linear approximations to x and y are

$$(68) \quad \begin{cases} x \cong X + Ha + M(b - b_o) + N(\gamma - \gamma_o) \\ y \cong Y + Pa + Q(B - b_o) + W(\gamma - \gamma_o) \ . \end{cases}$$

Since we want x = y = 0 when t = 0, no matter what values a, b, γ have, we conclude that H = M = N = P = Q = W = 0 when t = 0.

These are half of the initial conditions for Eq. (66). From Eq. (68) we obtain

$$(69) \qquad \begin{cases} \dot{x} = \dot{X} + \dot{H}a + \dot{M}(b - b_o) + \dot{N}(\gamma - \gamma_o) \\ \dot{y} = \dot{Y} + \dot{P}a + \dot{Q}(b - b_o) + \dot{W}(\gamma - \gamma_o) \end{cases}$$

Once again we let $t = 0$ and find

$$(70) \qquad \begin{cases} b \cos \gamma \cong b_o \cos \gamma_o + \dot{H}_o a + \dot{M}_o(b - b_o) + \dot{N}_o(\gamma - \gamma_o) \\ b \sin \gamma \cong b_o \sin \gamma_o + \dot{P}_o a + \dot{Q}_o(b - b_o) + \dot{W}_o(\gamma - \gamma_o) \end{cases}$$

from which we hope to get the initial values of $\dot{H}$, $\dot{M}$, etc. We may write

$$(71) \qquad b \cos \gamma = b \cos [\gamma_o + (\gamma - \gamma_o)]$$

$$= b [\cos \gamma_o \cos (\gamma - \gamma_o) - \sin \gamma_o \sin (\gamma - \gamma_o)] .$$

Now we are interested only in the case when $\gamma - \gamma_o$ is a small angle. Hence, we may say

$$(72) \qquad \begin{cases} \cos (\gamma - \gamma_o) \cong 1 \\ \sin (\gamma - \gamma_o) \cong \gamma - \gamma_o . \end{cases}$$

(These are the usual "small angle approximations.") Equation (71) now becomes

(73) $$b \cos \gamma \cong b \left[\cos \; _{o} - (\gamma - \gamma_o) \sin \gamma_o \right] .$$

Next we write $\underline{b}$ as $b_o + (b - b_o)$

(74) $$b \cos \gamma \cong \left[b_o + (b - b_o)\right] \left[\cos \gamma_o - (\gamma - \gamma_o) \sin \gamma_o \right]$$

$$\cong b_o \cos \gamma_o + (b - b_o) \cos \gamma_o - (\gamma - \gamma_o) b_o \sin \gamma_o$$

wherein we have discarded the term involving the <u>product</u> of the two <u>small</u> quantities $(b - b_o)$ and $(\gamma - \gamma_o)$. The result in Eq. (74), of course, is merely the linear approximation to $b \cos \gamma$ at (b_o, γ_o), and could have been obtained more easily perhaps by the methods discussed earlier. However, the technique used here is seen quite often, and is itself worthy of attention. A comparison now of Eq. (74) and Eq. (70) shows that $\dot{H} = 0$, $\dot{M} = \cos \gamma_o$, $\dot{N} = - b_o \sin \gamma_o$ when $t = 0$.

A similar treatment of $b \sin \gamma$ leads to

(75) $$b \sin \gamma \cong b_o \sin \gamma_o + (b - b_o) \sin \gamma_o + (\gamma - \gamma_o) b_o \cos \gamma_o$$

so that $\dot{P} = 0$, $\dot{Q} = \sin \gamma_o$, $\dot{W} = b_o \cos \gamma_o$ when $t = 0$. We now have all the initial conditions for Eq. (66).

It is a relatively routine matter to solve Eq. (66) and to evaluate the solutions at $t = T_o$. The results are

$$
(76)\quad \begin{cases} H(T_o) = -\, b_o^{\,2} \sin 2\gamma_o \displaystyle\int_o^{T_o/2} \sqrt{t^2 + \beta^2}\; dt \\[2ex] M(T_o) = \dfrac{b_o \sin 2\gamma_o}{g}\;;\; N(T_o) = -\,\dfrac{2b_o^{\,2} \sin^2 \gamma_o}{g} \\[2ex] P(T_o) = 2g^2 \displaystyle\int_o^{T_o/2} t^2 \sqrt{t^2 + \beta^2}\; dt \\[2ex] Q(T_o) = \dfrac{2b_o \sin^2 \gamma_o}{g}\;;\; W(T_o) = \dfrac{b_o^{\,2} \sin 2\gamma_o}{g}\;, \quad \text{where} \end{cases}
$$

$$
(77)\qquad \beta = \frac{b_o \cos \gamma_o}{g} \quad .
$$

After some tedious computations, the integrals in $H(T_o)$ and $P(T_o)$ can be evaluated, all the results can be substituted in Eq. (65), after which Eq. (65) can be solved for A, B, C, E, F, G, which in turn can be substituted in Eq. (56) to yield

$$
(78)\quad \begin{cases} T = T_o - \dfrac{b_o^{\,3} a}{4g^2 \sin \gamma_o} \left[(\cos^4 \gamma_o)\, \ell n \dfrac{1 + \sin \gamma_o}{\cos \gamma_o} \right. \\[2ex] \qquad \left. - \sin \gamma_o\, (2 - \cos^2 \gamma_o) \right] + \left(\dfrac{2 \sin \gamma_o}{g} \right) (b - b_o) \\[2ex] \qquad + \left(\dfrac{2 b_o \cos \gamma_o}{g} \right) (\gamma - \gamma_o) \end{cases}
$$

$$(78)\ \text{Cont.}\quad \begin{cases} R = R_o - \dfrac{b_o^4 a \cot \gamma_o}{4g^2}\left[(\cos^2 \gamma_o)(1 + 3\sin^2 \gamma_o)\, \ell n \dfrac{1 + \sin \gamma_o}{\cos \gamma_o}\right. \\ \qquad \left. - \sin \gamma_o (1 - 3\sin^2 \gamma_o)\right] + \left(\dfrac{2b_o \sin 2\gamma_o}{g}\right)(b - b_o) \\ \qquad + \left(\dfrac{2b_o \cos 2\gamma_o}{g}\right)(\gamma - \gamma_o)\ . \end{cases}$$

If $b_o = 305$ ft/sec, $\gamma_o = \pi/6$ radians, and $g = 32.2$ ft/sec^2, then Eq. (78) becomes

$$(79)\quad \begin{cases} T = 9.48 + 4320\, a + 0.0311\,(b - b_o) + 16.4\,(\gamma - \gamma_o) \\ R = 2500 \cdot 2160000\, a + 16.4\,(b - b_o) + 2890\,(\gamma - \gamma_o). \end{cases}$$

Thus a missile fired with a muzzle velocity of 305 ft/sec at an angle of elevation of 30° in a vacuum (a = o) would travel 2500 feet in 9.48 seconds. A missile fired in exactly the same way in an atomosphere for which a = 0.002 would travel minus 1820 feet in 18.12 seconds. This obviously ridiculous result should serve as a warning against the incautious use of results obtained by a perturbation analysis. The trouble here is that a = 0.002, while seemingly quite small, is not small enough. It is apparent that whether or not a quantity is small is a relative matter. The same quantity may be either small or large, depending on the use to which it is put.

A rough rule which can be used (though not always safely!) in situations like this is that the perturbation terms should amount to no more than a few per cent of the quantity being computed. Thus, in our problem, a should be restricted to about 5×10^{-5}, while $b - b_o$ could have an order of magnitude of about 10 ft/sec, and $\gamma - \gamma_o$ about 0.05 radians (about 3°). So long as a, b, γ have values consistent with these restrictions, Eq. (79) may be used to compute time of flight and range.

The problem just completed is a typical example of the troubles encountered in a perturbation analysis. Practically all of these troubles stem from the fact that the various functions are defined implicitly and cannot be obtained in an explicit analytic form. Consequently, it is necessary to employ implicit differentiation, which usually leads to rather formidable looking expressions. On the other hand, any equations which must be solved are linear. The perturbation method is well adapted to finding quick-but-not-too-dirty answers to a wide variety of practical and theoretical problems.

XII. AVERAGE OR MEAN

Whenever a series of measurements of a certain quantity (say the length of a stick, for example) is made, a collection of numbers x_r, $r = 1, 2, \ldots, N$, is obtained. Because of errors in the measuring instrument, and perhaps other factors beyond the control of the person making the measurements, the numbers x_r normally will not all be the same, but instead will vary to an extent depending on the factors affecting the measurement. The question then arises as to what value should be assigned to the measured quantity. Furthermore, it would be nice to have some estimate of whatever error is apt to be in the assigned value.

Let x represent the value to be assigned. Then $x - x_r$ is the residual or the deviation of the r^{th} measurement. A standard way to choose x is to choose it so that the function

$$E = E(x) = \sum_{r=1}^{N} (x - x_r)^2 \tag{80}$$

will take on its minimum value. E is merely the sum of the squares of the residuals (hence a positive quantity) and can be minimized by proper choice of x. By differentiating Eq. (80), setting the derivative equal to zero, and solving for x, we find

$$x = \bar{x} = \frac{1}{N} \sum_{r=1}^{N} x_r \,. \tag{81}$$

Thus the average of the x_r is the value which minimizes the sum of the squares of the residuals. It also has the following easily verified property.

$$\sum_{r=1}^{N} (\bar{x} - x_r) = 0 \quad . \tag{82}$$

XIII. VARIANCE

Although the average certainly does not display all the information contained in the complete set x_r, still, in one number, it gives us an important fact about that set. Another important number associated with the collection of the x_r is the variance which is defined by

$$\sigma^2 = \frac{1}{N} \sum_{r=1}^{N} (\bar{x} - x_r)^2 \tag{84}$$

and is merely the mean (or average) of the squares of the residuals, the residuals being taken with respect to the mean $\bar{x}$ of the x_r. The variance is a measure of how the measured values are spread around the mean. If the measurements are very accurate, so that the measured values are all close to the mean, then the residuals will all be small and the variance will also be small. However, if the measurements are not very accurate, so that the measured values sometimes depart widely from the mean, then the residuals will be large and the variance also will be large. Thus, it appears that the variance is related to the precision of the measurements and can serve as a measure of that precision.

These two statistical quantities, the mean and the variance, are the most important ones used in present day engineering design.

XIV. COMPUTATIONS BASED ON MEASURED VALUES

Very frequently it is necessary to compute a quantity on the basis of experimental data. For example, we may measure the side of a square and then compute the area of the square. Since there is an error in the measured value, then there will also be an error in the computed value. We may make a large number, N, of measurements of the side of the square and compute, for each one, the area of the square, obtaining thereby N values for the area. For each of these two sets of numbers, length of side and area, we may now compute the important statistical quantities, mean and variance. Clearly, there should be a relationship among these numbers.

More generally, consider the problem in which we measure x and compute y by $y = f(x)$. A collection of measurements x_r with mean $\bar{x}$ and variance σ_x^2 leads to a collection of computed values $y_r = f(x_r)$ with mean $\bar{y}$ and variance σ_y^2. The linear approximation to $f(x)$ at $x = \bar{x}$ is

(84) $$y \cong f(\bar{x}) + f'(\bar{x})(x - \bar{x}).$$ Hence

(85) $$y_r \cong f(\bar{x}) \pm f'(\bar{x})(x_r - \bar{x}) .$$

It follows from this that the mean of y is given by

(86) $$\bar{y} = \frac{1}{N} \sum_{r=1}^{N} y_r = \frac{1}{N} \sum_{r=1}^{N} [f(\bar{x}) + f'(\bar{x})(x_r - \bar{x})]$$

$$= f(\bar{x}).$$

Thus the average value of y can be obtained by computing it directly from the average of x. Equation (85) now may be written in the form

(87) $$y_r - \bar{y} \cong f'(\bar{x})\ (x_r - \bar{x})$$

and this can be used in turn to compute the variance of y.

(88) $$\sigma_y^2 = \frac{1}{N} \sum_{r=1}^{N} (y_r - \bar{y})^2 = \frac{1}{N} \sum_{r=1}^{N} [f'(\bar{x})]^2\ (x_r - \bar{x})^2$$

$$= [f'(\bar{x})]^2 \cdot \frac{1}{N} \sum_{r=1}^{N} (x_r - \bar{x})^2 = [f'(x)]^2\ \sigma_x^2 \ .$$

This shows how the variances of the measured and computed values are related.

It must be borne in mind that the results just derived are based on the use of a linear approximation and hence are subject to all the limitations implied by that approximation. Only in the case of fairly accurate measurements can we expect Eqs. (86) and (88) to be valid.

XV. FUNCTIONS OF MORE THAN ONE VARIABLE; COVARIANCE

More often than not, a computed value is based on measured values of several different quantities. Thus, let x_r, y_r, z_r be collections of measurements with means $\bar{x}$, $\bar{y}$, $\bar{z}$ and variances σ_x^2, σ_y^2, σ_z^2 respectively, and let $w = f(x,y,z)$. The linear approximation at $(\bar{x}, \bar{y}, \bar{z})$ is

(89) $$w \cong f(x, \bar{y}, \bar{z}) + A(x - \bar{x}) + B(y - \bar{y}) + C(z - \bar{z})$$

$$\left.\begin{aligned} A &= \frac{\partial f}{\partial x} \\ B &= \frac{\partial f}{\partial y} \\ C &= \frac{\partial f}{\partial z} \end{aligned}\right\} \quad \text{evaluated at} \quad \begin{aligned} x &= \bar{x} \\ y &= \bar{y} \\ z &= \bar{z} \end{aligned} \ .$$

It turns out, just as before, that $w = f(\bar{x}, \bar{y}, \bar{z})$. Thus, Eq. (89) may be written as

$$w_r - w \cong A(x_r - \bar{x}) + B(y_r - \bar{y}) + C(z_r - \bar{z}) \tag{90}$$

which leads to

$$\sigma_w^2 = \frac{1}{N} \sum_{r=1}^{N} [A(x_r - \bar{x}) + B(y_r - \bar{y}) + C(z_r - \bar{z})]^2 \tag{91}$$

$$= A^2 \sigma_x^2 + B^2 \sigma_y^2 + C^2 \sigma_z^2 + 2AB\, \sigma_{xy} + 2AC\, \sigma_{xz} + 2BC\, \sigma_{yz}$$

where σ_{xy}, σ_{xz}, σ_{yz} are new quantities defined by

$$\begin{cases} \sigma_{xy} = \frac{1}{N} \sum_{r=1}^{N} (x_r - \bar{x})(y_r - \bar{y}) \\ \sigma_{xz} = \frac{1}{N} \sum_{r=1}^{N} (x_r - \bar{x})\ (z_r - \bar{z}) \\ \sigma_{yz} = \frac{1}{N} \sum_{r=1}^{N} (y_r - \bar{y})(z_r - \bar{z})\ . \end{cases} \tag{92}$$

These new quantities are called the covariances of x and y, x and z, and y and z respectively. They give some idea of how the measurements of x, y, z are related. A very common situation is to have independent measurements, that is to say, the value obtained by measuring one quantity is not affected by, nor does it affect, the value obtained by measuring another quantity. Thus, if x and y are independent measurements, we may expect the product $(x_r - \bar{x})\ (y_r - \bar{y})$ to be

negative about as often as it is positive, and we may also expect its magnitude to be distributed fairly equally between the positive and negative values. The net result is that σ_{xy} should tend to zero as the number of measurements increase. In fact, we might take $\sigma_{xy} = 0$ as an indication that x and y are independent. If all the measurements are independent, then all the covariances vanish, and (91) reduces merely to

$$\sigma_w^2 = A^2 \sigma_x^2 + B^2 \sigma_y^2 + C^2 \sigma_z^2 \quad . \tag{93}$$

XVI. MATRIX FORM

A very convenient way to handle the variances and covariances is to write them in matrix form.

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix} \quad . \tag{94}$$

This is called the covariance matrix of x, y, z. Equation (91) now may be written as a matrix product.

$$\sigma_w^2 = (A\ B\ C) \begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} \quad . \tag{95}$$

There remains one obvious generalization. We may make measurements of n different quantities x, y, z, ..., and use these measurements to compute m other quantities u, v, w, ... Knowing the covariance matrix of x, y, z, ..., we would like to find the covariance matrix of u, v, w, ... Let

$$(96) \qquad M = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} & \cdots \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} & \cdots \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \quad n \times n \text{ matrix}$$

be the covariance matrix of x, y, z, ..., and

$$(97) \qquad \begin{pmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{uw} & \cdots \\ \sigma_{uv} & \sigma_v^2 & \sigma_{vw} & \cdots \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \quad m \times m \text{ matrix}$$

be the covariance matrix of u, v, w, ... Also let

$$(98) \qquad D = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} & \cdots \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} & \cdots \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} & \frac{\partial w}{\partial z} & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{pmatrix} \quad \begin{matrix} \text{m rows} \\ \\ \text{n columns} \end{matrix}$$

be the matrix of partial derivatives of u, v, w, ... with respect to x, y, z, ..., all these derivatives being evaluated at $(\bar{x}, \bar{y}, \bar{z}, \ldots)$. Then exactly the same sort of analysis as in the simpler cases shows that

$$(99) \qquad V = D\, M\, \tilde{D}$$

where $\tilde{D}$ is the transpose of D

XVII. STANDARD DEVIATION AND PRECISION

Suppose we have a large number of measurements x_r of a quantity x. Then the mean $\bar{x}$ and the variance σ^2 can be computed in the manner previously described. The number σ itself is called the standard deviation of the data. It is customary to express the precision of the measurements in terms of σ (rather than σ^2). In ordinary circumstances we may expect about 50 per cent of the measurements to fall in the interval from $\bar{x} - 0.67\sigma$ to $\bar{x} + 0.67\sigma$. (The coefficient

0.67 is determined on the basis of probability theory and certain quite reasonable assumptions about the nature of the errors to be expected in measurements. Different coefficients correspond to different percentages. Thus, about 90 per cent of the measurements fall in the interval from $\bar{x} - 1.65\sigma$ to $\bar{x} + 1.65\sigma$. It is in this sense that σ determines the precision of the set of measurements.)

XVIII. APPLICATION OF STATISTICS TO THE TRAJECTORY PROBLEM

To illustrate one of the uses to which these ideas from statistics may be put, we return to the trajectory problem and, in particular, to one of the results in Eq. (79).

$$R \cong 2500 - 2160000\, a + 16.4\,(b - b_o) + 2890\,(\gamma - \gamma_o) \tag{100}$$

where $b_o = 305$ ft/sec and $\gamma_o = \pi/6$ radians. To simplify the present discussion, we assume $a = 0$ (vacuum trajectory). Thus,

$$R - 2500 \cong 16.4\,(b - b_o) + 2890\,(\gamma - \gamma_o)\ . \tag{101}$$

This expresses the range error due to errors in muzzle velocity and angle of elevation. These errors are unavoidable. Consequently, we cannot expect to get a range of exactly 2500 feet on any given shot. A much more reasonable demand is to require, say, that 50 per cent of the shots should fall within 67 feet of the target, i.e., from 2433 to 2567 feet. In terms of standard deviation, this means that we want $0.67\sigma_R = 67$, or $\sigma_R^2 = 10^4$. But

$$\sigma_R^2 = (16.4)^2\, \sigma_b^2 + (2890)^2\, \sigma_\gamma^2\ . \tag{102}$$

(See Eq. (93). We assume, of course, that muzzle velocity and angle of elevation are independent.) This equation now constitutes a restriction on the values of σ_b^2 and σ^2, which are determined by the precision of manufacture of the missile and its launcher. Thus, if launchers are manufactured in such a way that $\sigma^2 = 10^{-4}$, then Eq. (102) gives $\sigma_b^2 = 34.1$. This means that the powder charge (or whatever controls the muzzle velocity) must be measured accurately enough so that $\sigma_b^2 = 34.1$ (or less) in order that at least the required percentage of hits will be made in the target area.

XIX. SOME APPLICATIONS OF PERTURBATION ANALYSIS TO SATELLITE PROBLEMS

Now that artificial earth satellites have become fact rather than fiction, there are many problems which no longer are of merely theoretical interest but are also of great practical importance. Information about the environment of the satellite can be collected and transmitted back to earth by instruments carried in the satellite. Besides this, a great deal more information can be derived from observations of the path type of information and ways of obtaining it.

The complexity of the problems precludes a detailed analysis in this brief report. Our intent here is to describe some problems, indicate how perturbation analysis is applied, and exhibit some of the resulting equations.

XX. GRAVITY

One important purpose of artificial earth satellites is to obtain information about the fine structure of the gravitational field of the earth. This can be done by observing the motion of a satellite which is far enough from the center of the earth so as not to be affected appreciably by the atmosphere, and yet close enough so that irregularities in the gravitational field are still large enough to cause observable variations in the satellite's trajectory. There is good reason to believe that the irregularities we are trying to detect will be "relatively small," so that perturbation techniques should provide a useful tool for finding the gross or "first order" effects to be expected.

XXI. TWO BODY PROBLEM:

We consider first a simplified problem (without the irregularities mentioned above) which we can solve analytically. This is the standard "two-body problem." We assume that the earth is a homogeneous sphere of mass m_E. As a result of this assumption, the force of attraction of the earth on an external body, say a satellite of mass m_S, is

$$F = \frac{G\, m_E\, m_S}{r^2} \tag{103}$$

where G is the gravitational constant and r is the distance from the center of the earth to the external body (assumed to be small enough, relative to the earth, to be taken as a point). Moreover, the force is directed toward the center of the earth. We take a rectangular coordinate system (x, y, z) whose origin is at the earth's center and whose axes have their directions fixed in inertial space. The components of the force F in the x, y, and z directions are $-\frac{Fx}{r}$, $-\frac{Fy}{r}$, and $-\frac{Fz}{r}$ respectively, since $\frac{x}{r}$, $\frac{y}{r}$, $\frac{z}{r}$ are the direction cosines for the force direction. Newton's Law, applied to the x-coordinate, gives

$$m_S \ddot{x} = -\frac{Fx}{r} = -\frac{G\, m_E\, m_S\, x}{r^3} \tag{104}$$

There are similar equations for the y and z coordinates. We divide by m_S and let $\alpha = G\, m_E$ to obtain the equations of motion of the satellite.

$$\begin{cases} \ddot{x} = -\dfrac{\alpha x}{r^3} \\ \ddot{y} = -\dfrac{\alpha y}{r^3} \\ \ddot{z} = -\dfrac{\alpha z}{r^3} \end{cases} \qquad r = \sqrt{x^2 + y^2 + z^2} \tag{105}$$

We let

$$(106) \qquad h = y\dot{z} - z\dot{y} \ , \quad b = z\dot{x} - x\dot{z} \ , \quad c = x\dot{y} - y\dot{x} \ .$$

Differentiation of these gives

$$(107) \qquad \frac{dh}{dt} = y\ddot{z} - z\ddot{y} = -\frac{\alpha yz}{r^3} + \frac{\alpha yz}{r^3} = 0, \quad \frac{db}{dt} = 0 = \frac{dc}{dt} \ .$$

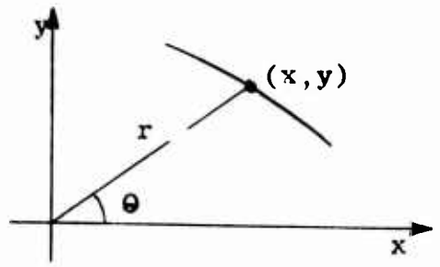
Hence h, b, c are constants. Furthermore,

$$(108) \qquad hx + by + cz = 0 \ ,$$

which shows that the satellite always lies in the plane with this equation. Inasmuch as h, b, c are constants, this plane does not alter its orientation with time.

XXII. POLAR FORM

Now that we know the satellite moves in a fixed plane, we move our coordinate system so that the x and y axes lie in that plane.



$$x = r \cos \theta$$
$$y = r \sin \theta$$

The equations of motion then become

$$\ddot{x} = -\frac{\alpha x}{r^3}\,, \quad \ddot{y} = -\frac{\alpha y}{r^3}\,, \quad z = 0. \tag{109}$$

It is convenient to introduce polar coordinates (r, Θ) by means of the usual transformations. Substitution in (109) leads to

$$\ddot{r} - r\dot{\Theta}^2 = -\frac{\alpha}{r^2} \tag{110}$$

$$r\ddot{\Theta} + 2\dot{r}\dot{\Theta} = 0 \quad .$$

The second of these equations can be written as

$$\frac{1}{r}\frac{d}{dt}(r^2\dot{\Theta}) = 0. \quad \text{Hence} \tag{111}$$

$$r^2\dot{\Theta} = n = \text{constant}. \tag{112}$$

We now can solve Eq. (112) for $\dot{\Theta}$ and substitute in the top of Eq. (110) to obtain a differential equation for r in terms of t.

$$\ddot{r} = \frac{n^2}{r^3} - \frac{\alpha}{r^2}\,. \tag{113}$$

Presumably this can be solved to obtain r as a function of t; however, it is much more informative to ignore the dependence of r upon t, and to find out how r depends on Θ. This will give us the <u>trajectory</u> of the satellite in the plane without specifying just where the satellite is in that <u>trajectory</u> at any given time. Later we can return to answer the question of time dependence. Hence, we differentiate $r = r(\Theta)$ twice with respect to t, after which some routine maneuvering with Eqs. (112) and (113) leads to

$$\frac{d^2r}{d\theta^2} - \frac{2}{r}\left(\frac{dr}{d\theta}\right)^2 = r - \frac{\alpha r^2}{n^2} \quad . \tag{114}$$

The change of variable $r = \frac{1}{\rho}$ now gives

$$\frac{d^2\rho}{d\theta^2} + \rho = \frac{\alpha}{n^2} \tag{115}$$

which is an easily solved linear differential equation. The solution is

$$\rho = \frac{\alpha}{n^2} + k \cos (\theta - \theta_o) \tag{116}$$

where k and θ_o are arbitrary constants which depend ultimately on the initial conditions.

From Eq. (116) we obtain

$$r = \frac{a(1-\epsilon^2)}{1 + \epsilon \cos (\theta - \theta_o)} \quad \text{where} \tag{117}$$

$$\epsilon = \frac{kn^2}{\alpha} \quad , \quad a = \frac{n^2}{a(1 - \epsilon^2)} \quad . \tag{118}$$

When $|\epsilon| \leq 1$, Eq. (117) is the polar equation of an ellipse with one focus at the pole. Thus, the trajectory of the satellite is an ellipse with the center of the earth at one focus. The size and shape of the ellipse, i.e., the semi-major axis a and the eccentricity ϵ depend on the initial conditions.

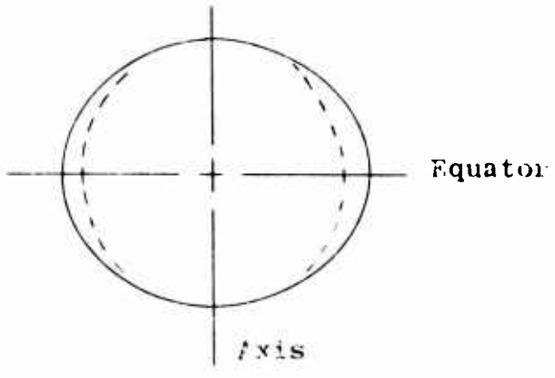
The hitherto neglected time dependence now can be attacked by substituting Eq. (117) into Eq. (112) to obtain a differential equation for θ in terms of t. We shall not do this since we do not need the results in what is to follow.

XXIII. ORBIT PARAMETERS

The general solution of Eq. (105) involves 6 arbitrary constants. To specify any particular trajectory and the position of the satellite in that trajectory, then, 6 conditions are required. A reasonable choice is the set of values x, y, z, $\dot{x}$, $\dot{y}$, $\dot{z}$ at some initial time t_o (usually taken to be $t_o = 0$)--the so-called initial conditions. The initial position and velocity coordinates thus constitute a set of 6 parameters which completely define the motion of the satellite. However, this is by no means the only set of 6 parameters which can be used. Other commonly used parameters are the 6 Keplerian parameters. Two of these specify the orientation of the orbital plane. Three more specify the orientation, size, and shape of the ellipse in the orbital plane. (These 3 are θ_o, a, ϵ.) The sixth parameter is the time at which the satellite passes through some particular point of its orbit, say the point closest to the focus (perigee). The particular problem being studied determines which set of parameters should be used.

XXIV. OBLATE EARTH

Unfortunately, instead of being a homogeneous sphere, the earth is a non-homogeneous, oblate spheroid. It is somewhat like a sphere with a belt around the equator.



The effect of the belt is to add a small component of force to the gravitational attraction of the sphere. The new force of attraction now has a component parallel to the axis of the earth, and consequently no longer is always directed toward the center of the earth. This brings with it all sorts of complications in the motion of the satellite.

The equations of motion of a satellite now become

$$(119)\quad \begin{cases} \ddot{x} = -\dfrac{\alpha x}{r^3} + \lambda\, x\, f \\ \ddot{y} = -\dfrac{\alpha y}{r^3} + \lambda\, y\, f \\ \ddot{z} = -\dfrac{\alpha z}{r^3} + \lambda\, z\, \left(f + \dfrac{3}{r^5}\right) \end{cases}$$

where $f = \dfrac{3(r^2 - 5z^2)}{2r^7}$ and λ is a parameter whose mangitude depends on the degree of oblateness of the earth. (For these equations, the x and y-axes must lie in the earth's equatorial plane while the z-axis points in the direction of the north pole. The form of the function f and the derivation of the above equations of motion are important matters which we shall not cover here.) The solutions of Eq. (119) now will depend on λ

$$(120)\quad x = x(t, \lambda)\ , \quad y = y(t, \lambda)\ , \quad z = z(t, \lambda)\ .$$

The linear approximations to these at $\lambda = 0$ are

$$(121)\quad x \cong X + M\lambda, \quad y \cong Y + N\lambda, \quad z \cong Z + P\lambda \quad \text{where}$$

$$(122)\quad \begin{cases} X = X(t) = x(t, o) \\ Y = Y(t) = y(t, o) \\ Z = Z(t) = z(t, o) \end{cases} \quad \text{and}$$

$$(123)\quad \begin{cases} M = \dfrac{\partial x}{\partial \lambda} \\ N = \dfrac{\partial y}{\partial \lambda} \\ P = \dfrac{\partial z}{\partial \lambda} \end{cases} \qquad \text{evaluated at } \lambda = 0.$$

Differentiation of Eq. (119) partially with respect to λ, followed by evaluation at $\lambda = 0$, gives the following differential equations for M, N, P.

$$(124)\quad \begin{cases} \ddot{M} = -\alpha\left[\dfrac{M}{R^3} - \dfrac{3X}{R^5}(XM + YN + ZP)\right] + XF \\ \ddot{N} = -\alpha\left[\dfrac{N}{R^3} - \dfrac{3Y}{R^5}(XM + YN + ZP)\right] + YF \\ \ddot{P} = -\alpha\left[\dfrac{P}{R^3} - \dfrac{3Z}{R^5}(XM + YN + ZP)\right] + Z\left(F + \dfrac{3}{R^5}\right) \end{cases}$$

where $R = \sqrt{X^2 + Y^2 + Z^2}$ and $F = \dfrac{3(R^2 - 5Z^2)}{2R^7}$. Now X, Y, Z presumably are known functions of t (although they may be rather complicated) since, when $\lambda = 0$, the equations of motion reduce to the case which we previously managed to solve analytically to obtain the Kepler ellipse. Hence, the system of simultaneous linear differential equations in Eq. (124) can be solved and the results substituted in Eq. (121) to find the actual trajectory of the satellite. This would indeed be a formidable undertaking on paper. Fortunately, we shall be able to make use of the equations in Eq. (124) without actually solving them.

XXV. ROTATION OF ORBITAL PLANE

We return to Eq. (106), namely

$$(106)\quad h = y\dot{z} - z\dot{y}, \quad b = z\dot{x} - x\dot{z}, \quad c = x\dot{y} - y\dot{x}.$$

Previously, when $\lambda = 0$, we found that h, b, c were constants. When $\lambda \neq 0$, however, there is no reason to suppose that they will still be constants. Hence we write

$$(125) \quad h = h(t, \lambda), \quad b = b(t, \lambda), \quad c = c(t, \lambda)$$

and introduce the linear approximations

$$(126) \quad \begin{cases} h \cong h(t, o) + H\,\lambda \\ b \cong b(t, o) + B\,\lambda \\ c \cong c(t, o) + C\,\lambda \end{cases} \qquad \text{where}$$

$$(127) \quad \begin{cases} H = \dfrac{\partial h}{\partial \lambda} \\ B = \dfrac{\partial b}{\partial \lambda} \\ C = \dfrac{\partial c}{\partial \lambda} \end{cases} \qquad \text{evaluated at } \lambda = 0.$$

The equation

$$(108) \quad hx + by + cz = 0$$

still holds, even if h, b, c are not constants. When $\lambda = 0$, Eq. (108) is the equation of the orbital plane, and h, b, c (which are constant) determine the orientation of that plane. When $\lambda \neq 0$, Eq. (108) is still the equation of a plane, and h, b, c still determine the orientation; but, since h, b, c now vary with time, the plane is no longer fixed. We may still refer to it as the orbital plane, since the satellite always lies in it, but we must permit it to move around. Our present purpose is to find out how it moves. This we hope do to by studying the derivatives H, B, C.

By differentiating Eq. (106) partially with respect to λ and then setting $\lambda = 0$, we find

$$(128)\qquad \begin{cases} H = Y\dot{P} + \dot{Z}N - Z\dot{N} - \dot{Y}P \\ B = Z\dot{M} + \dot{X}P - X\dot{P} - \dot{Z}M \\ C = X\dot{N} + \dot{Y}M - Y\dot{M} - \dot{X}N \end{cases}$$

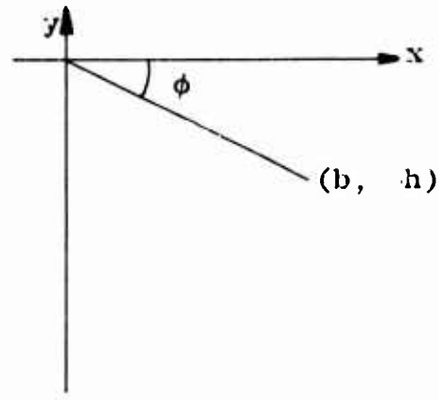
which expresses H, B, C in terms of the known functions X, Y, Z and the solutions M, N, P of Eq. (124). Differentiation of Eq. (128) with respect to time yields

$$(129)\qquad \begin{cases} \dot{H} = Y\ddot{P} - \dot{Y}P + Z\ddot{N} - \ddot{X}N \\ \dot{B} = X\ddot{M} - \ddot{Z}M \quad \ddot{P} - \ddot{X}P \\ \dot{C} = X\ddot{N} - \ddot{X}N + Y\ddot{M} - \ddot{Y}M \,. \end{cases}$$

In these, the second derivatives of M, N, P can be replaced through use of Eq. (124), while the second derivatives of X, Y, Z are given by $\ddot{X} = -\alpha X R^{-3}$ etc. This simplifies Eq. (129) to

$$(130)\qquad \dot{H} = \frac{3YZ}{R^5}\,, \quad \dot{B} = -\frac{3XZ}{R^5}\,, \quad \dot{C} = 0\,.$$

It is a simple matter to verify that the point (b, -h, o) lies in the orbital plane as well as in the xy-plane. Hence, the line segment joining this point to the origin lies in the intersection of these two planes. Let ϕ be the angle between the line segment and the x-axis. Then



$$(131) \qquad \phi = - \arctan \frac{h}{b}$$

is one of the two keplerian parameters determining the orientation of the orbital plane. (The other is the angle of inclination of the orbital plane to the equatorial plane.)

$$(132) \qquad \phi = \phi(t, \lambda) \cong \phi(t, o) + \Phi\lambda, \quad \Phi = \left(\frac{\partial \phi}{\partial \lambda}\right)_{\lambda = 0}$$

From Eq. (131) we find

$$(133) \qquad \frac{\partial \phi}{\partial \lambda} = - \frac{b \frac{\partial h}{\partial \lambda} - h \frac{\partial b}{\partial \lambda}}{b^2 + h^2}$$

which, upon setting $\lambda = 0$, gives

$$(134) \qquad \Phi = - \frac{b_o H - h_o B}{b_o^2 + h_o^2} \quad . \qquad \text{Hence}$$

$$(135) \qquad \dot{\Phi} = - \frac{b_o \dot{H} - h_o \dot{B}}{b_o^2 + h_o^2} = - \frac{3Z}{R^5} \cdot \frac{h_o X + b_o Y}{b_o^2 + h_o^2} = \frac{3c_o Z^2}{(b_o^2 + h_o^2) R^5}$$

since $h_o X + b_o Y + c_o Z = 0$ from Eq. (108). Therefore, differentiation of Eq. (132) with respect to t gives

$$\dot{\phi} = \frac{3\lambda\, c_o\, Z^2}{(b_o^2 + h_o^2)\, R^5} \quad . \tag{136}$$

This simple expression contains a large amount of information, only part of which we shall endeavor to extract. The variable part of $\dot{\phi}$ is Z^2R^{-5}, which is never negative. Hence, $\dot{\phi}$ never changes sign. This means that ϕ is either always increasing or always decreasing, depending on the sign of c_o. This, in turn, means that the line segment which defines ϕ rotates in the xy-plane (though not uniformly) and the orbital plane rotates with it. The instantaneous rate of rotation is given by Eq. (136). Although we shall not do it here, it is possible without much trouble to compute the average rate of rotation of the orbital plane from Eq. (136). This average rate is easily observable in actual satellites, and, since it is directly proportional to λ, can be used to estimate the value of λ. This, of course, gives a measure of the earth's oblateness. In this way, observation of the motion of artificial satellites gives us information about the shape of the earth.

The preceding long-winded exposition was designed to show how perturbation techniques have been applied to find the effects of the earth's oblateness on the orbits of satellites. The particular effect which was studied was the rotation of the orbital plane. There are other effects also, all of which have been studied by the same method and are to be found discussed in various published sources.

XXVI. SATELLITE TRACKING

Although it is both interesting and important to have the type of information given by the above analysis, there are even more important factors to be considered in the practical problem of tracking a satellite. To "track" a satellite means to determine its position (and perhaps velocity) in space, and also involves the ability to predict future

positions on the basis of present knowledge. The obvious way to do this is to integrate Eq. (119), namely

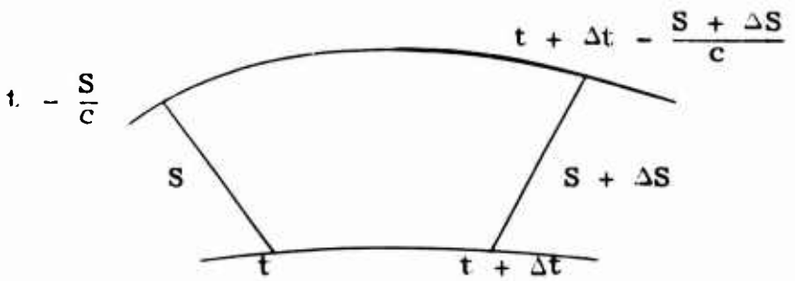
$$(119)\qquad \begin{cases} \ddot{x} = -\dfrac{\alpha x}{r^3} + \lambda\, x\, f \\ \ddot{y} = -\dfrac{\alpha y}{r^3} + \lambda\, y\, f \\ \ddot{z} = -\dfrac{\alpha z}{r^3} + \lambda\, x\, (f + \dfrac{3}{r^5}) \end{cases} \qquad f = \frac{3(r^2 - 5z^2)}{2r^7} \qquad r = \sqrt{x^2 + y^2 + z^2}\ .$$

A high-speed digital computer can handle this task both easily and satisfactorily. It is only necessary to supply the machine with initial values of position and velocity, i.e., x_o, y_o, z_o, $\dot{x}_o$, $\dot{y}_o$, $\dot{z}_o$, and with a time interval Δt, after which the machine produces values of x, y, z, $\dot{x}$, $\dot{y}$, $\dot{z}$ at the times $t_o + n\,\Delta t$, $n = 1, 2, 3, \ldots$. This all sounds very simple; unfortunately, the big problem still remains. It is still necessary to make the machine computations apply to some particular satellite which we wish to track. This means that we must properly choose the six parameters (or initial conditions) x_o, y_o, z_o, $\dot{x}_o$, $\dot{y}_o$, $\dot{z}_o$. Mathematically, then, the tracking problem reduces to the determination of 6 parameters.

In order to determine the parameters we must, by observing the satellite, make measurements of some quantity which depends on the parameters. Then we can compare the measured values with computed values of the quantity and attempt to juggle the parameters so as to bring the computed values into agreement with the measured ones. Naturally there must be some mathematical technique for doing this juggling. The one which is used is the method of least squares.

XXVII. DOPPLER SHIFT

Several different things have been measured in various tracking schemes. The emphasis at APL is on the doppler shift.



Let the upper curve in the diagram be the trajectory of a satellite carrying a radio transmitter operating at a frequency of f_T cycles per second. Let the lower curve be the path of an observer on the earth. The observer moves relative to the satellite path because of the rotation of the earth on its axis. At time t, the observer receiver a signal from the satellite which was emitted at time $t - \frac{S}{c}$. S is the distance from the observer to the position occupied by the satellite when the signal was emitted, and c is the speed of light. At a time Δt seconds later, when satellite and observer have both moved to new positions, the situation is as pictured in the diagram. The time interval over which the satellite has been transmitting its signal is $(t + \Delta t - \frac{S + \Delta S}{c}) - (t - \frac{S}{c}) = \Delta t - \frac{\Delta S}{c}$ seconds, so that the total number of cycles emitted is $f_T(\Delta t - \frac{\Delta S}{c})$. All of these cycles are received by the observer in time Δt. Hence, the observer thinks the transmitter is operating at a frequency of $f_o = \frac{f_T(\Delta t - \frac{\Delta S}{c})}{t} = f_T (1 - \frac{1}{c}\frac{\Delta S}{\Delta t})$. The amount of the doppler shift is the difference between the apparent frequency and the actual frequency, i.e., $f_o - f_T = -\frac{f_T}{c}\frac{\Delta S}{\Delta t}$. The instantaneous doppler shift $D = D(t)$ now can be obtained by letting $\Delta t \to 0$.

$$D = -\frac{f_T}{c}\frac{dS}{dt} \qquad (137)$$

It is fairly easy to make very precise measurements of the function D. To be able to compute D, we must be able to compute $\frac{dS}{dt}$. This derivative is the rate of change of the distance between satellite and observer. Presuming that we know the position of the observer, we will be able to compute S (and hence $\frac{dS}{dt}$) if we have a way for computing the position of the satellite. But this is just what the integration of Eq. (119) gives us, namely the position of the satellite as a function of the 6 parameters. Hence, D itself is a function of those same 6 parameters. The route may be devious and the expressions complicated (fortunately not too complicated), but the important fact remains that D depends on the parameters we wish to determine.

XXVIII. PERTURBATION OF THE TRAJECTORY

As indicated previously, the method of least squares is used to find the parameters. All the details of this procedure are to be found in several APL reports. Here is it merely our intention to show how perturbation methods become involved in the computations, and to indicate the nature of the resulting expressions. Basically our problem is to find out how the satellite trajectory (which determines D) depends on the parameters. The dependence, of course, is given implicitly in the original differential equations of motion, Eq. (119). What we need for use with the least squares procedure is something more explicit. The usual situation is that we have a rough idea of what the parameters should be, and we are trying to get more precise values. This is the same as saying that we know approximately what the trajectory should be, and that we are trying to find a better trajectory which presumably is "close" to the estimated trajectory. Hence, we attempt to find the new trajectory by perturbing the old.

Each of the trajectory variables x, y, z is a function of the 6 parameters x_o, y_o, z_o, $\dot{x}_o$, $\dot{y}_o$, $\dot{z}_o$. Each may be differentiated partially with respect to each parameter. There are 18 such derivatives. We, of course, are interested in these derivatives when they have been evaluated at $x_o = x_{oo}$, $y_o = y_{oo}$, etc. Let

$$(138)\qquad \begin{cases} Q = \dfrac{\partial x}{\partial x_o} \\ W = \dfrac{\partial y}{\partial x_o} \\ U = \dfrac{\partial z}{\partial x_o} \end{cases} \qquad \text{evaluated at } x_o = x_{oo}, \text{ etc.}$$

These are three of the 18 derivatives. (The first zero subscript on x_{oo} means that it is an initial condition; the second subscript denotes a particular set of initial conditions.) The linear approximations to x, y, z at $x_{oo}, y_{oo}, z_{oo}, \dot{x}_{oo}, \dot{y}_{oo}, \dot{z}_{oo}$ are

$$(139)\qquad \begin{cases} x = x^* + Q\,(x_o - x_{oo}) + \\ y = y^* + W\,(x_o - x_{oo}) + \\ z = z^* + U\,(x_o - x_{oo}) + \end{cases} \qquad \text{terms involving the other 15 partial derivatives}$$

where x^*, y^*, z^* are the trajectory coordinates computed for the special case $x_o = x_{oo}$, etc. We shall content outselves with finding just one of the 18 simulataneous equations which can be solved to obtain the 18 derivatives.

Differentiate the first equation in Eq. (119) partially with respect to x_o.

$$(140)\qquad \frac{d^2}{dt^2}\left(\frac{\partial x}{\partial x_o}\right) = -\alpha\left\{\frac{1}{r^3}\frac{\partial x}{\partial x_o} - \frac{3x}{r^5}\left[x\frac{\partial x}{\partial x_o} + y\frac{\partial y}{\partial x_o} + z\frac{\partial z}{\partial x_o}\right]\right\}$$
$$+ \lambda f \frac{\partial x}{\partial x_o} + \lambda x \left\{\frac{\partial f}{\partial x}\frac{\partial x}{\partial x_o} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x_o} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x_o}\right\}.$$

Now evaluate at $x_o = x_{oo}$, etc.

$$(141) \qquad \frac{d^2Q}{dt^2} = -\alpha\left\{\frac{Q}{r^3} - \frac{3x^*}{r^5}\left[x^* Q + y^* W + z^* U\right]\right\} + \lambda f^* Q + \lambda x^* \left[f_x^* Q + f_y^* W + f_z^* U\right]$$

where $f^* = \frac{3\left[r^2 - 5(z^*)^2\right]}{2r^7}$ and now $r = \sqrt{(x^*)^2 + (y^*)^2 + (z^*)^2}$.

Also $f_x^* = \left(\frac{\partial f}{\partial x}\right)$ evaluated at $x = x^*$, $y = y^*$, $z = z^*$. There are 17 more equations similar to this one which can be solved for Q, W, U, etc. This sounds like a nearly impossible task, but fortunately, in terms of the capabilities of modern electronic digital computers, it is a rather simple one.

Once Q, W, U, etc. are obtained, then Eq. (139) is the desired explicit representation of the trajectory in terms of the parameters. With this we conclude this illustration of some of the problems involved in satellite tracking.

Each time a satellite passes near a receiving station, a measurement of the doppler shift can be made. By using several receiving stations and making measurements over long periods of time (weeks, months, or even years), a large amount of doppler data can be accumulated. It would be eminently reasonable to suppose that out of all this data we should be able to extract very precise estimates of the orbit parameters, the precision being limited only by the unavoidable "noise" in the data. We would expect the orbit parameters to approach constant values; however, this is not at all the case! The parameters computed from the data of one week will be different from those computed from the data of the preceding week. From week to week the parameter values will change or drift. This peculiar behavior is caused by errors in the values of α and λ used in the equations of motion, Eq. (119). The number α, for example, is proportional to the mass of the earth, which is known to an accuracy of only about 1 part in 10,000. The number λ is known even less accurately. On the other hand, doppler shift measurements can be made with accuracies up to 1 part in 10^8, which is several orders of magnitude better than our knowledge of α and λ. Of course, the satellite is not afflicted by this lack of knowledge. Its motion is governed by the exact values of α and λ, whereas we poor mortals are compelled to use more or less inaccurate estimates. This means that

the computed doppler shift will inherit the errors in α and λ and transmit those errors to the computed values of the orbit parameters.

Since the drift of the orbit parameters is directly related to α and λ, then we should be able to get improved estimates of α and λ by measuring the amount of drift. This is one of the results to be expected from satellite tracking. Here again, perturbation analysis is an appropriate technique for handling this complex problem.

There is one more set of parameters which we wish to consider. This is the set of 3 parameters which define the location of the receiving station, e.g., longitude, latitude, and distance from the center of the earth. Heretofore, we have considered these to be known. By using receiving stations of known position, we can eventually, as indicated above, very accurately track a satellite. As pointed out previously, this means that we can accurately predict its future positions. If now we make doppler measurements at a receiving station whose position is not known very well, then we can employ exactly the same techniques used in tracking to determine the 3 parameters of the receiving station. Of course, this problem is simpler since only 3 parameters are involved. This is known as satellite navigation. It is potentially a very accurate navigation scheme. In this particular application, too, perturbation analysis eases the labor required in the computations. It is to be especially noted that an adequate satellite navigation scheme depends primarily on a good tracking scheme.

XXIX. BRIEF COMMENTS ON PERTURBATION ANALYSIS

In common with many other useful procedures, the ideas behind perturbation analysis are quite simple. The application of these ideas to particular problems leads to work of varying degrees of difficulty. The principal problem, of course, is the evaluation of derivatives. The difficulties arise because it is usually necessary to resort to implicit differentiation to obtain the desired derivatives. Furthermore, there may be several intermediate variables to be considered; that is to say, the variables we are after may be related implicitly to certain other variables which, in turn, are related implicitly to the basic variables appearing in the mathematical formulation of the problem. Most of the problems in satellite work are of this type. As partial compensation for these troubles, perturbation analysis leads to linear equations.

APPENDIX: ELEMENTARY ASPECTS OF MATRIX ALGEBRA

A matrix is a rectangular array of numbers.

$$\begin{pmatrix} 3 & 0 & -10 & 1 \\ -1 & 5 & 4 & 0 \\ 7 & 2 & -2 & -5 \end{pmatrix}$$

The above example is a matrix with 3 rows and 4 columns, hence a total of 12 elements. Two matrices are equal if and only if they have the same shape, i.e., same number of rows and same number of columns, and have their elements in the same order. The following three matrices are not equal.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 2 \\ 3 & 4 & 5 \end{pmatrix}$$

The purpose of a matrix is to permit a collection of numbers to be handled as an entity. The ability to do so turns out to be extremely useful.

Addition---The sum of two matrices is defined if and only if the matrices have the same shape. In that case, the sum is obtained by adding corresponding elements together.

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix} + \begin{pmatrix} 5 & 2 & 1 \\ -6 & 2 & -5 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 4 \\ -3 & 6 & 0 \end{pmatrix}$$

Multiplication---If A and B are matrices, then the product AB is defined only under the following condition: The number of columns in the left hand factor A must be the same as the number of rows in the right hand factor B. When this condition is satisfied, then the elements of the product AB can be computed by a procedure which sounds complicated but is quite easy to carry out. The element in the m^{th} row and n^{th} column of AB is obtained by multiplying the elements of the m^{th} row of A by the corresponding elements of the n^{th} column of B (first with the first, second with the second, etc.) and adding the products thus obtained.

$$A = \begin{pmatrix} -2 & 1 \\ 3 & 4 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}$$

$$AB = \begin{pmatrix} -2 & 1 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} -2+3 & -4+4 & -6+5 \\ 3+12 & 6+16 & 9+20 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & -1 \\ 15 & 22 & 29 \end{pmatrix} .$$

BA is not defined. It will be noted that the product of two matrices has the same number of rows as the left factor and the same number of columns as the right factor.

Vectors---A matrix with only one column is usually called a column vector. A matrix with only one row is usually called a row vector.

Linear Equations---Simultaneous linear equations can easily be written in a matric form.

$$3x + 2y = 1$$
$$7x + 5y = 3$$

Let $C = \begin{pmatrix} 3 & 2 \\ 7 & 5 \end{pmatrix}$ be the matrix consisting of the coefficients of the unknowns.

Let $X = \begin{pmatrix} x \\ y \end{pmatrix}$ be the column vector whose elements are the unknowns. Let $K = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ be the column vector whose elements are the constant terms in the equation. Then

$$\begin{pmatrix} 3 & 2 \\ 7 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad \text{or} \quad CX = K.$$

Unit Matrix---A square matrix is one which has the same number of rows as columns. This number is called the order of the square matrix. A very special square matrix is the unit matrix I whose elements are all zero except for those on the main diagonal, all of which are one.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$ second order

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$ fourth order

Sometimes a subscript is used to designate the order, e.g., I_4. If A is any matrix and I is a unit matrix of the appropriate order, then AI = A and IA = A.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 1+0 & 2+0 & 3+0 \\ 0+3 & 0+4 & 0+5 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1+0+0 & 0+2+0 & 0+0+3 \\ 3+0+0 & 0+4+0 & 0+0+5 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}$$

<u>Inverse of a Matrix</u>---Let $C = \begin{pmatrix} 3 & 2 \\ 7 & 5 \end{pmatrix}$ as before, and let $B = \begin{pmatrix} 5 & -2 \\ -7 & 3 \end{pmatrix}$.

Then $BC = \begin{pmatrix} 5 & -2 \\ -7 & 3 \end{pmatrix}\begin{pmatrix} 3 & 2 \\ 7 & 5 \end{pmatrix} = \begin{pmatrix} 15-14 & 10-10 \\ -21+21 & -14+15 \end{pmatrix}$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I.$$

Similarly, $CB = I$. It is apparent that the matrix B has a special relationship to the matrix C. B is said to be the inverse of C, and it written $B = C^{-1}$. Of course, it would be equally proper to say that C is the inverse of B, i.e., $C = B^{-1}$. This leads to the statement $(C^{-1})^{-1} = C$.

In general, whenever two square matrices H and M (of the same order) satisfy the condition $HM = MH = I$, then each is the inverse of the other. Thus,

$$H = \begin{pmatrix} 4 & 1 & 1 \\ 7 & 2 & 2 \\ 11 & 2 & 3 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -8 & 3 & 1 \end{pmatrix}$$

are inverses of each other since

$$HM = \begin{pmatrix} 8+1-8 & -4+1+3 & 0-1+1 \\ 14+2-16 & -7+2+6 & 0-2+2 \\ 22+2-24 & -11+2+9 & 0-2+3 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$$

and (similarly) $MH = I$.

Not all (square) matrices have inverses. For example, $A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ has no inverse. To show this, let $B = \begin{pmatrix} w & x \\ y & z \end{pmatrix}$ be an arbitrary matrix of order two. Then

$$AB = \begin{pmatrix} 0 & 0 \\ y & z \end{pmatrix}$$

and it is clear that AB can never be equal to $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ no matter what values are given to w, x, y, z. Matrices <u>without</u> inverses are said to be <u>singular</u>; matrices <u>with</u> inverses are <u>nonsingular</u>. The <u>determinant</u> of a square matrix is the determinant whose elements are the matrix elements in exactly the same order. The condition for a matrix to be nonsingular is that its determinant be different from zero.

The problem of finding the inverse of a matrix is a very important one; however, it will not be discussed here.

The importance of a matrix inverse perhaps can be estimated by considering the following example. The set of linear equations

$$\begin{aligned} 4x + y + z &= 5 \\ 7x + 2y + 2z &= -1 \\ 11x + 2y + 3z &= 3 \end{aligned}$$

can be written in the matrix form as $HX = K$, where $X = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $K = \begin{pmatrix} 5 \\ -1 \\ 3 \end{pmatrix}$, and $H = \begin{pmatrix} 4 & 1 & 1 \\ 7 & 2 & 2 \\ 11 & 2 & 3 \end{pmatrix}$ is the matrix considered previously with the inverse $H^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -8 & 3 & 1 \end{pmatrix}$. Multiply both sides of $HX = K$ by H^{-1} to obtain

$$H^{-1} HX = IX = X = H^{-1} K$$

which says that the solution of the matrix equation is

$$X = H^{-1} K = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 1 & -1 \\ -8 & 3 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 1 \\ -40 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Hence $x = 11$, $y = 1$, $z = -40$ is the solution of the original set of equations. This is easily verified by substitution or even by solving the equations in some other way.

Transpose, Symmetric Matrices---The transpose of a matrix A is the matrix obtained by interchanging rows and columns. We shall denote the transpose of A by $\tilde{A}$.

$$A = \begin{pmatrix} -2 & 1 \\ 3 & 4 \end{pmatrix} \qquad \tilde{A} = \begin{pmatrix} -2 & 3 \\ 1 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix} \qquad \tilde{B} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 5 \end{pmatrix}$$

The transpose of a row vector is a column vector, and conversely. A matrix which is equal to its own transpose is said to be symmetric.

$$S = \begin{pmatrix} 1 & -4 & 3 \\ -4 & 5 & 1 \\ 3 & 1 & 2 \end{pmatrix} = \tilde{S} \quad \text{is symmetric.}$$

REFERENCES

1. Modern Mathematical Methods and Models, Vol 1, Mathematical Association of America, 1958. (Section on linear approximation.)

2. Modern Mathematics for the Engineer, University of California Engineering Extension Series, edited by E. F. Beckenbach, McGraw-Hill, 1956. (Many topics, including perturbation methods.)

3. Engineering Cybernetics, by H. S. Tsien, McGraw-Hill, 1954. (Ch. IJ: Control Design by Perturbation Theory.)

4. Mathematics for Exterior Ballistics, by Gilbert Ames Bliss, John Wiley & Sons, 1944.

5. New Methods in Exterior Ballistics, by Forest Ray Moulton, University of Chicago Press, 1926.

6. "Perturbation Calculus in Missile Ballistics," R Dronick, Journal of the Franklin Institute, Vol. 251, 4 April 1951.

7. "Perturbation Analysis of Near Earth Satellites," by R. E. Roberson, Journal of the Franklin Institute, Vol. 264, 1957, p. 181 and p. 269.

8. Perturbations on the Orbits of Near Earth Satellites, by R. R. Newton, APL Internal Report, 2 June 1958.

9. Second Order Perturbations on the Orbits of Near Earth Satellites, by R. R. Newton, APL Internal Report, 20 June 1958.

10. Theoretical Analysis of Doppler Radio Signals from Earth Satellites, by W. H. Guier and G. C. Weiffenbach, APL Bumblebee 276, April 1958.

11. Variance Equations for Use in the Doppler Tracking of Satellites, APL BBD-573, 9 February 1959.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

ADAPTIVE METHODS AND DEVICES

by

A. G. Carlton

I. INTRODUCTION

In considering adaptive methods and devices, it would be convenient to know what is meant by the term adaptive. If one attempts to distinguish the adaptive from the non-adaptive, he may have great difficulty in determining the dividing line, and more trouble in getting agreement with his conclusions. We consider one method or device more adaptive than another, to the extent that it produces satisfactory results for a significantly greater variety of conditions.

Some people refer to a device as adaptive if it in some way detects and compensates for inadequacies in its operation. Others consider such a device as demonstrating active adaptation and refer to devices without this feature, but operating satisfactorily over a wide range of conditions, as having passive adaptation. Some people seem to consider a device adaptive only if its operation cannot be understood; we see no merit in such views.

Evolution has tended to produce living organisms which are much more adaptive than the products of human technology. Our understanding of the adaptive principles used in nature is far too meager to yield easy application to engineering designs. Many able people are attempting to discover the methods used by central nervous systems, but the problems are still very formidable.

In discussing adaptive methods, we shall consider four topics: 1) adaptive features of standard techniques; 2) adaptive filters, designed to handle a variety of input signal and noise characteristics; 3) compensation for unknown transfer functions; 4) compensating for noise correlated, in an unknown way, with the remaining input.

II. ADAPTIVE FEATURES OF STANDARD METHODS

Before considering complex methods designed to increase adaptivity, it would be well to notice adaptive features of some methods very widely used but not generally considered particularly adaptive. These are the closed loop, the matched filter, and the bang-bang servo.

Closed Loop Methods---These methods are used in order to obtain variety and complexity in living organisms, in many aspects

of activity. One of the achievements of cybernetics was its recognition of the powerful use of closed loop methods in life processes. Technology made little use of closed loop methods until nearly the present day, and even now the closed loop is foreign to the thought of most people, even scientists and engineers.

The first clear cases I can find of closed loop methods are in computation - Newton's methods of successive approximation for finding a square root of a number or a zero of a polynomial.

These computation methods illustrate some interesting features of closed loop systems. Consider computing the square root of 2. There is a classical method of doing this, involving repeated processes similar to long division but more complex, and if one mistake is made the answer may be grossly wrong. Newton's simple method involves repeated long division, is easily remembered, and is virtually foolproof. The idea is to guess a square root, divide it into the number, and average the divisor and quotient for the next approximation. Thus, I guess 1, divide it into 2, and get 2. The next guess is $\frac{1}{2}(1 + 2) = 1.5$, giving 1.3333 as quotient, 1.416 for the next guess. At the next stage one gets 1.41421. Repeating the process would give 12-figure accuracy. An error at any step but a final unchecked one would not affect the final result, only lengthening the computation. This tendency to correct errors is characteristic of closed loop methods, as is the precision obtainable with simple operations, and to some extent the nulling of the difference between output and input. Since the square root can be determined even by a man who forgets the classical method, the technique also shows that lower quality components can be used than in the open loop method.

A closed loop method of finding the root of a polynomial $f(x) = 0$ is to select a trial x_o, evaluate $f(x_o)/f'(x_o)$, and choose as the next approximation $x_1 = x_o - f(x_o)/f'(x_o)$. When this method works, it is far easier to get precise results than with formal algebraic methods, particularly if the polynomial is of more than the fourth degree. Sometimes the method fails, either because $f'(x_o)$ is zero or because successive approximations diverge. The tendency of instability of the process can usually be alleviated by suitably taking account of $f''(x)$, which is analogous to the use of lead or derivative terms for stabilizing servo systems. The classical servo system was Watt's steam engine governor, first analyzed mathematically

by Clark Maxwell about one hundred years ago, leading to Routh's study of stability of servo systems.

The simplest and most widely studied closed loop system is the servo or inverse feedback system, sketched in Fig. 1A.

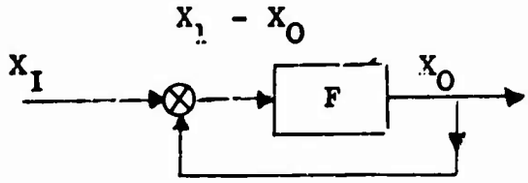


Fig. 1A

Fig. 1B

The output is produced by the operator F, a linear transfer function, operating on the "error signal" $X_I - X_O$. Solving the relation

$$X_O = F(X_I - X_O)$$

for the transfer function X_O/X_I, we obtain

$$\frac{X_O}{X_I} = \frac{F}{F + 1} \quad .$$

From this, it is hard to see any advantage for the closed loop system. If for some real frequency ω, $F(\omega)$, the system is obviously unstable; and if the system is stable it is equivalent to an open loop system with transfer function $F/(F + 1)$ as shown in Fig. 1B.

The closed loop system can be made stable by proper design, however, and may be very superior to the open loop system if we consider departure from the ideal as in Fig. 1C or Fig. 1D.

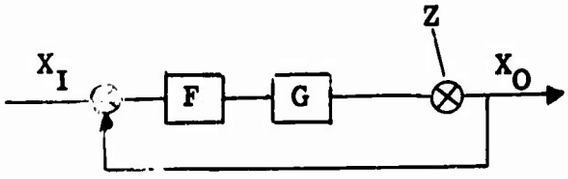


Fig. 1C

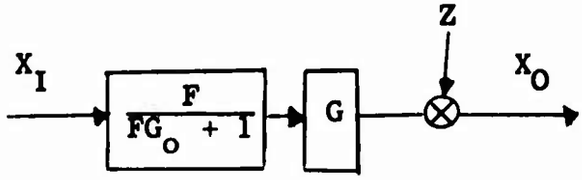


Fig. 1D

Assuming G a linear transfer function not known and fixed in advance, we find in the closed loop system

$$\left.\begin{aligned} X_O &= Z + FG(X_I - X_O) \\ \frac{X_O}{X_I} &= \frac{FG + Z/X_I}{FG + 1} \end{aligned}\right\} \quad \text{(closed loop)}$$

$$X_O = \frac{FG}{FG + 1} X_I + \frac{Z}{FG + 1} \quad \text{(closed loop), while}$$

$$X_O = \frac{FG}{FG_O + 1} X_I + Z \quad \text{(open loop)} \quad .$$

Variations in G and Z affect the output directly in the open loop system, much less strongly in the closed loop (negligibly for FG sufficiently large). When the operator G is non-linear, producing distortion, the results are similar. The possibility for instability of the closed loop is increased by the variations we have considered, but in general with proper design the transfer function from input to output can be made much less dependent on these variations than is that of an open loop system. We note also that the open loop is badly unstable if Z is produced by a constant $Z \neq 0$, whereas the closed loop can be made stable against such disturbances. For any reasonable definition of "satisfactory performance" the closed loop system is satisfactory over a much wider range of permissible G and Z.

A linear filter has an adaptive feature if satisfactory performance is defined in terms of mean square deviations, since such performance depends only on the first and second

moments of the signal and noise, and is thus unaffected by any further details of the probability distribution.

A bang-bang servo system is one in which there is a so-called bistable element, sometimes referred to as a switch or relay. This element produces the maximum possible numerical output with the same sign as the input. In a missile autopilot, the output might be control surface position or rate. In such a system the gain may be very large, but positive instability is prevented by the effective gain of the bistable element being reduced as the system nears instability. Such systems have been studied extensively in recent years at APL and elsewhere. A practical problem with these systems is the excessive dissipation of power.

The super-regenerative detector has some points of similarity with the bang-bang servo, and demonstrates some vaguely similar features.

III. ADAPTIVE FILTERS

You have learned something of the problem of filtering signal from noise, and have learned about Wiener's filtering theory, giving minimum mean square error in equilibrium, for stationary signal and noise with known spectral densities. We now consider the problem of devising a filter to give satisfactory performance for a variety of input characteristics.

Polynomial Signal---Suppose that the signal is a polynomial in time, of degree n, but is not otherwise restricted. Then the filter must follow such a polynomial perfectly to avoid unbounded errors. Writing the signal as

$$x(t) = c_0 + \frac{c_1}{1!}\,t + \frac{c_2}{2!}\,t^2 + \ldots + \frac{c_n}{n!}\,t^n \quad ,$$

we see that

$$x = c_0 + \frac{c_1}{p} + \frac{c_2}{p^2} + \ldots + \frac{c_n}{p^n}$$

$$= \frac{c_n + c_{n-1}\,p + \ldots + c_1 p^{n-1} + c_0 p^n}{p^n} \quad .$$

The steady state error in following x is

$$[x - x\,F(p)]_{p=0} = [x\{1 - F(p)\}]_{p=0} .$$

A necessary and sufficient condition for zero steady state error is that

$$\lim_{p\to 0} \frac{1 - F(p)}{p^n} = 0 .$$

Writing F in the form

$$F = \frac{a_0 + a_1 p + a_2 p^2 + \ldots}{b_0 + b_1 p + b_2 p^2 + \ldots} ,$$

we must have $a_0 = b_0$, $a_1 = b_1$, ..., $a_n = b_n$, with

$$1 - F = \frac{\left(b_{n+1} - a_{n+1}\right) p^{n+1} + \left(b_{n+2} - a_{n+2}\right) p^{n+2} + \ldots.}{b_0 + b_1 p + b_2 p^2 + \ldots.} .$$

If the filtering is accomplished with a servo, the forward transfer function G of the servo is

$$\frac{F}{1-F} = \frac{a_0 + a_1 p + a_2 p^2 + \ldots.}{\left(b_{n+1} - a_{n+1}\right) p^{n+1} + \left(b_{n+2} - a_{n+2}\right) p^{n+2} + \ldots} .$$

The forward transfer function in a servo filter must integrate once more than the degree of the polynomial.

If the noise is stationary with known spectral density $\Theta(\omega)$, one can select the coefficients of F subject to the restriction that $(a_0, a_1, \ldots, a_n) = (b_0, b_1, \ldots b_n)$, which make the mean square error, $\int |F(\omega)|^2 \Theta(\omega) d\omega$, as small as desired, at the expense of large transient errors.

The problem of designing a time-varying filter without transient or steady-state error in following a polynomial of given degree, and minimizing the mean square error due to stationary noise, has been solved theoretically. The solution of this problem in practice appears to be unduly difficult, and perhaps of little value in view of the artificiality of the assumed signal characteristics.

The theory can be easily extended to consider signals consisting of a stationary component of spectral density $\Phi(\omega)$ plus any polynomial of degree n. The optimal constant-coefficient filter is that which minimizes

$$\int \left\{ |F|^2 \Theta + |1 - F|^2 \Phi \right\} d\omega$$

by choice of F among those transfer functions satisfying the restrictions $(a_0, a_1, \ldots a_n) = (b_0, b_1, \ldots b_n)$. We note that the filter described by Hanson is optimal not only for the noise and signal spectral densities assumed, but also for cases in which the signal has any linear function of time as an additional component.

The Minimax Filter---The minimax principle, acting to minimize the maximum expected loss, has many advantages, whether one is dealing with alert enemies or intractable nature. The approach just mentioned is minimax for signals which may include any n^{th} degree polynomial, but have a remainder with known spectral density.

A useful class of signals to consider is that in which the spectral density $\Phi(\omega)$ is not known, but subject to linear constraint of the form

$$\int \Phi(\omega) \alpha(\omega) \, d\omega \leq 1 \, .$$

For example, if the mean square signal acceleration is limited, we would have $\alpha(\omega)$ proportional to ω^4; for limited mean square signal velocity, $\alpha(\omega)$ is proportional to ω^2.

By use of Lagrange multipliers and calculus of variations it is found that the minimax filter satisfies

$$\left|1 - F_{\Phi^*,\Theta}\right|^2 = \lambda\alpha(\omega) \qquad \omega : \Phi^* > 0$$

$$\leq \lambda\alpha(\omega) \qquad \omega : \Phi^* = 0,$$

with $F^*_{\Phi^*,\Theta}$ the optimal filter for Φ^* and Θ, Φ^* the maximum signal spectral density, and λ chosen to satisfy the constraint.

At this point it may appear difficult to determine the minimax filter, as the solution given is far from explicit. For simple forms of noise density $\Theta(\omega)$, however, the solution is easy. For white noise, $\Theta(\omega)$ constant, one can show that

$$\left|1 - F_{\Phi^*,\Theta}\right|^2 = \frac{\Theta}{\Phi^* + \Theta} .$$

To illustrate the procedure, consider $\alpha = \omega^4/a^2$, indicating limited mean square target acceleration. Equating our expressions for $\left|1 - F_{\Phi^*,\Theta}\right|^2$, and setting $a^2/\lambda \equiv \omega_o^4$,

$$\frac{\Theta}{\Phi^* + \Theta} = \frac{\omega^4}{\omega_o^4} , \qquad \omega : \Phi^*(\omega) > 0$$

$$\leq \frac{\omega^4}{\omega_o^4} , \qquad \omega : \Phi^*(\omega) = 0$$

giving

$$\Phi^* = \begin{cases} Q\left(\dfrac{\omega_o^4}{\omega^4} - 1\right), & \omega^2 < \omega_o^2 \\ 0, & \omega^2 > \omega_o^2 \end{cases}.$$

ω_o is determined by the constraint

$$1 = \int \Phi(\omega)\ \alpha(\omega) d = \int_{-\omega_o}^{\omega_o} Q\left(\frac{\omega_o^4}{\omega^4} - 1\right) \frac{\omega^4 d\omega}{\alpha^2} = \frac{2Q}{\alpha^2} \int_{\omega_o}^{\omega_o} \left(\omega_o^4 - \omega^4\right) d\omega,$$

from which $\omega_o = \left(\dfrac{5}{8} \dfrac{a^2}{Q}\right)^{1/5}$.

The minimax F can only be approximated in practice. A very good approximation is similar to the filter against accelerating targets, mentioned by Hanson, but with a damping factor of $\sqrt{3/8}$ rather than $\sqrt{1/2}$. That is,

$$F = \frac{1 + \sqrt{3/2}\ p/\omega_o}{1 + \sqrt{3/2}\ p/\omega_o + p^2/\omega_o^2}.$$

Estimating Input Spectral Densities---Perhaps the most obvious method for coping with a variety of input spectral densities is to process the input to estimate the spectral densities and choose the filter on the basis of these estimates. There are considerable difficulties associated with this approach, beginning with the problem of estimating two spectral densities on the basis of one time series, and suitably changing filter characteristics based on these estimates. With this method there is no check on the correctness of the adjustment, which is open loop. The final problem is that in radar tracking and missile guidance the input is typically not available - geometry provides feedback and only the so-called error signal is measured.

Adjusting Error Spectral Densities---A key to closed loop adjustment of the filter characteristics is the relation, stated above, that for white noise,

$$\left| 1 - F_{\Phi,\Theta} \right|^2 = \frac{\Theta}{\Theta + \Phi} .$$

Since the transfer function between input and error signal is $\frac{X_I - X_o}{X_I} = 1 - F(\omega)$, this implies that if the filtering is optimal, the spectral density of the error signal is

$$(\Theta + \Phi) \left| 1 - F_{\Phi,\Theta} \right|^2 = \Theta .$$

Closed loop adjustment of the filter is thus possible by changing filter characteristics in such a way as to maintain the error signal spectrum approximately flat. In practical applications of this concept, the error signal energy in a low pass band has been compared with that in a higher pass band. A filter of standard form $F = \frac{1 + 2\zeta \tau p}{1 + 2\zeta \tau p + \tau^2 p^2}$ has been used, with τ varied at a rate depending on the ratio of energies. This simple scheme has resulted in excellent adaptation to varying amounts of signal and noise, resulting in near optimal filtering.

Optimal Filtering in Transient and Non-Stationary Cases---By exploiting the principles of least squares, it is possible in many cases to design a filter which minimizes mean square errors in the transient response as well as in equilibrium, and which can cope with nonstationary inputs. This extension from filtering which is optimal only in steady state and requires stationary signal and noise certainly increases the adaptivity.

One would like to apply to this type of filtering some of the adaptive methods mentioned above - minimax principles,

estimation of input spectra, and closed loop adjustment based on the spectrum of error signals. Only the first two of these principles have been applied. Minimax design is considerably more difficult than in the case of the Wiener optimal filter, but has been accomplished by Busey in at least one problem. The basic method of adaptation is estimation, by sampling, of the spectra of the noise and of signal driving functions; in some applications the raw material has been available. Simplified closed loop filter adjustment based on the error signal spectrum does not appear possible. One could use measured deviations of the error signal spectral density to adjust estimates of strength of noise and signal driving components, thus using closed loop adjustment, but it is by no means clear that this would be superior to direct estimation.

Limiting RMS Output Acceleration---The RMS output acceleration of a missile must be limited. A system designed by Follin to adapt the filtering to limit RMS acceleration due to noise and avoid the most harmful effects of hitting acceleration limits is described in Reference (1), pp. 20-21.

IV. COMPENSATING FOR UNKNOWN TRANSFER FUNCTIONS

It is frequently the case, particularly in autopilot design, that a variable, not completely known transfer function is a significant part of the servo loop. In autopilot design, one may attempt to compensate by varying gains as functions of ram or static pressure, but there will often be large residual variations not properly compensated. The bang-bang systems mentioned earlier have often been suggested for autopilots to cope with these variations. In practice, however, the autopilot is typically compensated as well as feasible for pressure variations, and parameters selected to yield compromise performance over the region of possible variation. We shall now consider some special techniques for cases in which such compromise is inadequate.

Maximizing Stable Regulator Gain---The first problem considered is to maximize regulator gain without instability. This problem arose in attempting to control roll precisely in a missile with enormous variation in aerodynamic gain. With standard techniques, the roll autopilot gain had to be kept low to avoid instability with high aerodynamic gain; roll control was inadequate with low aerodynamic gain. The solution depended on the fact that the resonant frequency was

nearly unstable, the aerodynamic changes being primarily in gain. Energy in this frequency region was detected, and compared with energy calculated for neutral stability. The difference was used to vary the autopilot gain. The difference in performance of the system, according to whether the energy is measured before or after the variable gain, is significant. Both can be made stable and are correct in steady state, but the closed loop system has zero velocity lag and transient response independent of the critical gain.

Sensitivity Feedback---One of the least desirable types of autopilot, from the theoretical standpoint, is a wing position control system, which attempts to make wing deflection proportional to demanded acceleration. Such autopilots have been used, however, either because of their great simplicity or because the alternative accelerometer feedback system causes trouble due to amplification of a missile body vibration and bending. Satisfactory operation of a wing position control system requires adequate knowledge of the ratio of acceleration to wing position. A solution proposed by T. W. Sheppard was to compare average acceleration with average demanded acceleration, and use this comparison to adjust the gain of the autopilot. This approach was found to be feasible even in some quite complicated cases. It was also found, the hard way, that sufficiently shoddy components for changing the gain could invalidate any performance analysis. There is still a requirement for compact, reliable, reasonably precise gain-changing devices for use in guided missiles.

Tracer Signals---In the laboratory, a common method for determining system response is to insert signals at the input and observe the output. The use of this method in flight with autopilots could furnish useful information on airframe and other responses. Tracer signals of various sorts have been proposed for various purposes, but are usually rejected because of interference with flight or strong dependence on details of dead space, friction, etc., when too small or too high frequency to disrupt normal missile operation.

Tracer signals can be used effectively in many systems which function intermittently. As an example, they can facilitate adjustment of gains of pulse radar IF strips.

REFERENCES

1. "Recent Developments in Fixed and Adaptive Filtering," by A. G. Carlton and J. W. Follin, Jr., presented at Venice, Italy, AGARD Seminar, 1956.

2. "Optimal Filtering in Missile Guidance," by A. G. Carlton, presented at Indianapolis Meeting, AAAS, 28 December 1957; see also sixth paper in this compilation.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

RECENT DEVELOPMENTS IN FIXED AND ADAPTIVE FILTERING

by

A. G. Carlton and J. W. Follin, Jr.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

ABSTRACT

Classical optimal filtering methods have been extended to a large class of problems in which the input has incompletely specified characteristics. By minimax principles the optimal filter and the best input are determined. Two problems of time-varying filters are considered, first the optimal settling of filters to steady state and second the design of adaptive filters which adjust to varying or unknown environment.

I. INTRODUCTION

Linear filtering theory is largely based on the fundamental work of Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," 1950, which in many respects paralleled the independent work of Kolmogorov.

It may be well to review the problem considered by Wiener. He assumed that there was available the entire past history of a time series consisting of signal plus noise possibly correlated with the signal, that all processes involved were stationary and indeed ergodic to the second order, with known auto-correlation and cross-correlation functions. He wished to determine the realizable linear filter to apply to the signal in order to minimize the mean square difference between the output and the message translated by an assigned positive or negative time interval. Wiener solved this problem by using variational analysis on the weighting function to obtain an integral equation, then by using subtle Fourier analysis to solve the integral equation. Multiple time series were handled by an extension of this technique.

It will be noted that the problem solved by Wiener contains two restrictions beyond the assumptions: first, optimization is restricted to linear functions; second, the loss function whose expectation is minimized is the squared error.

In attempting to extend filtering theory, it is appropriate to modify or eliminate various of these assumptions or restrictions. Various investigators have widened the field of permissible filters and dealt with alternative loss functions. Nonstationary processes have been considered; some trivial obvious results have been obtained, and some adaptive filters appear suitable, but little has been done that is both significant and rigorous. In this paper we shall remove the assumption of complete knowledge of the correlation functions and also indicate some minor extensions of the basic theory and techniques, consider the optimum filter with only portions of the signal history available and attempt to classify the types of adaptive filters.

II. THE FREQUENCY APPROACH

The reader of Wiener's work will note that although his basic problem is formulated in terms of time series,

correlation functions, and weighting functions, his solutions are expressed in terms of spectral densities and transfer functions. It appears reasonable, consequently, to set up the problem in the frequency domain. From the elementary properties of spectral densities, we have

$$(2.1) \qquad \sigma^2 = \int \left\{ \left| e^{i\omega\alpha} - F \right|^2 M + \left| F \right|^2 N \right\} \, d\omega \, .$$

where σ^2 is the mean square error, $F(\omega)$ the filter transfer function, α the time translation, and M and N the signal noise spectral densities, so normalized that the signal power is $\int M(\omega) \, d\omega$. All integrals are taken over the entire real frequency axis unless otherwise indicated, and the dependence of the variables on ω will usually not be indicated. It is assumed here and henceforth that the signal and noise are independent; this entails no loss in generality in the classical developments, where M can be regarded as the sum of the signal and signal on noise spectral densities, N as the sum of the noise and noise on signal spectral densities.

It is useful to consider the spectral densities as resolutions of the signal power into a continuum of frequency components. From this standpoint it is clear that ergodic properties are not relevant to the problem of linear filtering, although the optimal filter may be nonlinear if the second-order characteristics are not ergodic.

Before applying variations to minimize σ^2 by choice of F, let us indicate some extensions of this relation to problems beyond the original one. In the first place it will be noted that Eq. (2.1) is valid even though the power of signal, noise, or both is unbounded; it is not necessary that the correlation function exist. It will be seen that this may be of importance.

A trivial generalization of Eq. (2.1) is to replace $e^{i\omega\alpha}$ by the Fourier transform $Y(\omega)$ of any desired linear operator on the signal, e.g. by $i\omega$ for the derivative. Thus,

$$(2.2) \qquad \sigma^2 = \int \left\{ \left| Y - F \right|^2 M + \left| F \right|^2 N \right\} \, d\omega .$$

Another easy extension is to the case in which it is desired to weight errors unequally for various frequency components. A symmetric non-negative function $W(\omega)$, so normalized that $\int W(\omega)\, d\omega = 1$, could be inserted to obtain

$$\sigma_*^2 = \int \left\{ \left| Y - F \right|^2 M + \left| F \right|^2 N \right\} W \, d\omega \,, \tag{2.3}$$

which is of course formally equivalent to Eq. (2.2) with M and N replaced by MW and NW.

The next extension is to minimize σ^2 subject to a restriction on the mean power of the output or some other linear function of the output. As an example, if σ^2 must be minimized subject to the restriction that the output acceleration power must be less than β, i.e., that

$$\int (M + N) \left| F \right|^2 \omega^4 \, d\omega \leq \beta \,, \tag{2.4}$$

we should minimize

$$\sigma^2 + \lambda\beta = \int \left\{ \left| Y - F \right|^2 M + \left| F \right|^2 N + \lambda \, (M + N) \left| F \right|^2 \omega^4 \right\} d\omega, \tag{2.5}$$

and select λ to satisfy Eq. (2.4). In certain cases, especially with non-gaussian processes, restraints such as Eq. (2.4) may preferably be applied only to the noise. In this case the integral to be minimized can be reduced to Eq. (2.2) by suitable definition of N. This type of side condition can be introduced formally as here, or can be used in the definition of the class over which F is optimized. Several simultaneous side conditions can be introduced in the same way.

The final extension to be considered is to replace the class of relizable transfer functions by other classes, say $\mathcal{F}$, as appropriate.

III. MINIMIZATION OF σ^2 BY $F \subset \mathcal{F}$

We now consider the selection of F within the class $\mathcal{F}$ to minimize σ^2. An increment J or F will produce an incremental σ^2 of

$$\sigma_J^2 + F - \sigma_F^2 = \int |J|^2 (M + N)\, d\omega + \int \bar{J} \left[(M + N) F - MY \right] d\omega, \tag{3.1}$$

a relation obtained by noting that every transfer function has an even real part and an odd imaginary part.

The transfer function F will be optimal in $\mathcal{F}$ if the right-hand integral of Eq. (3.1) is non-negative for every J such that $J + F \subset \mathcal{F}$ and $\int |J|^2 (M + N)\, d\omega$ is finite.

The absolutely optimal F is, from Eq. (3.1), evidently

$$F_o = \frac{MY}{M + N} . \tag{3.2}$$

The optimal realizable F is given by

$$F_R = \frac{1}{(M + N)^+} \left[\frac{MY}{(M + N)^-} \right]_+ , \tag{3.3}$$

where the new symbols denote factorization and decomposition of a meromorphic function as

$$H \simeq H^+ H^- + H_+ + H_- , \tag{3.4}$$

H^+ being analytic and without poles or zeroes in the lower half plane, H^- the conjugate of H^+; H_+ and H_- have no poles in the lower half plane and upper half plane, respectively. Polynomial terms of H appear in H_+. If $M + N$ is not factorable, $F_R = F_O$.

To check the validity of Eq. (3.3), substitute it in Eq. (3.1), obtaining

$$(3.5)\quad \sigma_J^2 + F_R - \sigma_{F_R}^2 = \int |J|^2 (M + N)\, d\omega + \int \bar{J}\, (M + N)^- \left[\frac{MY}{(M + N)^-}\right]_- d\omega;$$

the latter integral is zero, by contour integration over the upper half plane, for J sufficiently convergent, i.e., for $\int |J|^2 (M + N)\, d\omega$ finite.

As an example let us consider the spectra

$$(3.6)\quad M = \Theta/\omega^4 \quad , \quad N = \phi \quad , \quad \text{with } Y = 1 \quad ,$$

the absolute optimum (realizable with infinite delay) is then

$$(3.7)\quad F_O(p) = \frac{1}{1 + \phi\, p^4/\Theta}$$

and the optimal realizable F is

$$(3.8) \qquad F_R(p) = \frac{1 + \sqrt{2}\ p\ (\phi/\Theta)^{1/4}}{1 + \sqrt{2}\ p\ (\phi/\Theta)^{1/4} + p^2\ (\phi/\Theta)^{1/2}} \quad .$$

This transfer function is the zero velocity lag loop with 0.7 critical damping which will be discussed below. The corresponding errors are

$$(3.9) \qquad \sigma_o^2 = .3535\ \phi^{3/4}\ \Theta^{1/4}$$

$$\sigma_R^2 = 1.414\ \phi^{3/4}\ \Theta^{1/4} \quad .$$

A very useful relation can be obtained by manipulation of the second integral of Eq. (3.1) as follows:

$$(3.10) \qquad \int \bar{J}\ [(M + N)\ F - MY]\ d\omega = \int \bar{J}/\bar{F}\ [(M + N) \left| F \right|^2 - MY\bar{F}]\ d\omega$$

$$= \int \bar{J}/\bar{F}\ [(M + N) \left| F \right|^2 - \mathrm{Re}MY\bar{F} - \mathrm{Im}\ MY\bar{F}]\ d\omega \ .$$

This integral is zero for sufficiently convergent realizable J if

$$(3.11) \qquad \left| F_R \right|^2 (M + N) = \mathrm{Re}\ MY\bar{F} + \mathrm{Re{:}Im}\ MY\bar{F} \ ,$$

where Re:ImH represents the real complement to the given imaginary function such as to render H realizable.

This can be more formally expressed by the Bode relation

(3.12) $\left|F_R\right|^2 (M + N) = \frac{1}{i\pi} \int_* \frac{w\,(MY\overline{F})_w\, dw}{\omega^2 - w^2}$, where $\int_*$ indicates disregard of poles at $w = \pm \omega$, provided MY does not diverge for large ω.

Solutions of Eq. (3.11) for simple forms of MY are readily obtained; for example,

(3.13) $$\left|F_R\right|^2 (M + N) = \begin{cases} MY, \text{ if MY is constant} \\ MY\overline{F}\,(ia), \text{ if } MY = \frac{C \cdot a^2}{a^2 + \omega^2} \\ MY + k, \text{ if } MY = c_o + c_2 \omega^2 + c_4 \omega^4 \end{cases}$$

$$\left[k \text{ determined by } \int \log \left|F_R\right|^2 d\omega = 0\right] .$$

By symmetry under the interchange of $M \leftrightarrow N$ and $F \leftrightarrow (Y - F)$ we obtain

(3.14) $\left|Y - F_R\right|^2 (M + N) = NY$, if NY is constant.

For ordinary filtering, with $Y = 1$, these define the optimal spectral density of the error signal in a servo type filter, and can be used to construct an adaptive filter which will become optimal for any signal spectrum.

Expressions for $\left|F_R\right|^2$, based on Eq. (3.11) are quite useful in optimizing filters with side conditions such as limited mean square output, since the expressions for $\left|F_R\right|^2$ are frequently more convenient to use than those for F_R in determining the Lagrange multipliers.

IV. MINIMAX FILTERING

Let us consider now some typically statistical problems, in which one has incomplete knowledge of the spectrum of noise, of signal, or both. We discuss below adaptive quasi-linear filters which appear suitable for cases in which one ectrum is completely unknown, and which can cope with cases volving a spectral density of known form but unknown magni-ude.

The problems considered by Wiener were essentially probabilistic, i.e., the system is completely described in terms of appropriate probability measures; the problems we are now considering are statistical, in that we are dealing with a system defined by probabilities, some of which are unknown. Our problem in this case is one of statistical estimation of a function of the signal. Our estimating function should be optimal in some sense. One of the most logical criteria for an estimate, developed by Abraham Wald, is that it minimizes the maximum expected loss; that is, each filter is assessed on the basis of the expected loss with the possible system which is least favorable for the given filter, and the optimal filter is that filter for which the maximum expected loss is minimum. This formulation of statistical decision theory is very similar to two-person game theory, independently developed by von Neuman. We adopt this criterion and consider as optimal the minimax filter, with the loss function proportional to the squared error. In a completely prescribed system, the minimax linear filter is the Wiener optimal linear filter.

Minimax theory offers a strong justification for the use of linear filters. If the distribution functions of the processes are not known but the class of possible distributions includes Gaussian distributions, the minimax filter is linear, since with a linear filter the mean square error is independent of the form of the distribution function, and with a non-linear filter the mean square error exceeds that with a linear filter when the processes are Gaussian.

Typical problems encountered in practice involve situations in which the noise process is known to be limited in power, in mean square velocity, etc.:

$$(4.1) \qquad \int N \, d\omega = C_0, \int N \, \omega^2 \, d\omega = C_1, \text{ etc.}$$

Such restrictions can be put in the general form

$$(4.2) \qquad \int N\,\Theta\,d\omega = 1,$$

where Θ is a prescribed symmetric non-negative function, K a prescribed positive number. It can be shown straightforwardly that the variation in σ^2 due to variation n in N is a non-negative function of n, zero if and only if n = 0, plus

$$(4.3) \qquad \int n \left[\left|F_{M,N}\right|^2 - \lambda\,\Theta \right] d\omega,$$

where $F_{M,N}$ is the optimal transfer function with M and N, and λ is a Lagrange multiplier to be selected to satisfy Eq. (4.2). From Eq. (4.3) and the fact that n + N must be non-negative, it follows that the maximum N is N_o given by

$$(4.4) \qquad \left|F_{M,N_o}\right|^2 = \lambda\,\Theta \qquad \omega:\ N_o > 0$$

$$\leq \lambda\,\Theta \qquad \omega:\ N_o = 0.$$

This result is easily generalized to the case of several inequality restrictions

$$(4.5) \qquad K_j \int N\Theta_j\,d\omega = 1 \qquad (j = 1,2\ldots,\ k)$$

with the K_j not specified but ≥ 1.

The maximin N is N_o satisfying

$$(4.6) \qquad \left|F_{M,N_o}\right|^2 = \sum \lambda_j K_j \Theta_j \qquad \omega: \ N_o > 0$$

$$\leq \sum \lambda_j K_j \Theta_j \qquad \omega: \ N_o = 0$$

with the λ_j, K_j satisfying Eq. (4.5) and also

$$(4.7) \qquad \lambda_j (K_j - 1) = 0, \qquad (j = 1,2\ldots, k) .$$

Similar results are obtained for cases in which the signal spectral density is subject to one or more equalities or inequalities, and where both spectra are limited only by such restrictions.

The results just given have derived the maximin spectrum or spectra, but our object is to determine the minimax filter. For this purpose we now prove that

$$\min_F \max_N \sigma^2_{N,F} = \max_N \min_F \sigma^2_{N,F}$$

ant that the minimax F is F_{M,N_o}. To prove this, we observe that

$$\min_F \max_N \sigma^2_{N,F} \leq \max_N \sigma^2_{N,F_{M,N_o}} =$$

$$\max_N \int \left\{ N \left|F_{M\ N_o}\right|^2 + M \left|Y - F_{M,N_o}\right|^2 \right. d\omega$$

$$= \sum \lambda_j + \int M \left|Y - F_{M,N_o}\right|^2 d\omega = \max_N \min_F \sigma^2_{N,F} .$$

We have thus shown that

$$(4.8) \qquad \min_{F} \max_{N} \sigma^2_{N,F} \leq \max_{N} \min_{F} \sigma^2_{N,F} .$$

But by the fundamental theorem of game theory,

$$(4.9) \qquad \min_{F} \max_{N} \sigma^2_{N,F} \geq \max_{N} \min_{F} \sigma^2_{N,F} .$$

from which it follows at once that the "game" is determined, with the minimax filter being F_{M,N_o} and the maximin N being the N_o previously defined.

The basic result of the minimax approach to optimum filtering is that the errors depend in the second order on the spectra and the form of the filter. As a consequence if a suitable approximation to the optimal form is used and the parameters are adjusted properly the resulting system will be satisfactory.

V. TIME VARYING FILTERS

Let us now consider the problem of filtering when only a finite and perhaps fragmentary history of the signals is available. In this case the filter parameters are variable, and we must assume a particular form of transfer function. In general the steady state optimal filter with variable band-pass and damping is best. We may attack this problem in the time domain by considering the rate of change of the filtering errors and adjusting parameters to maximize the rate of decrease of the error.

For a linear system the tracking accuracy may be described in terms of the variances of the tracking error and error rate. As an example let us consider the simple zero velocity lag feedback system in Fig. 1. The input

consisting of signal and noise $x(t) + x_n(t)$ is at the left, the output, x_c, at the right. The equations of the system are

$$\frac{dx_c}{dt} = \dot{x}_c + b\,(x - x_c) + b\,x_n \tag{5.1}$$

$$\frac{dx_{\dot{c}}}{dt} = a\,(x - x_c) + a\,x_n$$

Note, in explanation of the notation, that $x_{\dot{c}} \neq \dot{x}_c$.

Actually the principal interest centers upon the errors $\epsilon_c = x - x_c$, $\epsilon_{\dot{c}} = \dot{x} - \dot{x}_{\dot{c}}$. In the second diagram is shown the error loop equivalent to the original signal loop. The signal x now appears as an acceleration input to the first integrator. This is a significant advantage when, as is often the case, the acceleration spectrum of the signal is known.

So far a, b are unrestricted. If $\ddot{x}$ and x_n both have flat spectral densities Θ, ϕ respectively then this filter is optimal with the values of a and b previously derived. The present purpose is to extend the optimization to the transient period. The gains a, b in this case are time functions and the resulting system, while not necessarily optimum among all possible systems, is the best obtainable with a given structure. The key to the solution lies in setting up the differential equations relating the variances and covariances of the integrator outputs ϵ_c, $\epsilon_{\dot{c}}$. Write Eq. (5.1) in terms of the errors and express the solution in the neighborhood of t as a power series in Δt. Terms beyond the first degree in Δt are not required. The result is most easily obtained by direct use of the second figure

$$\epsilon_c = \epsilon_{c_o} + \epsilon_{\dot{c}_o}\,\Delta t - b\,\epsilon_{c_o}\,\Delta t + b\int_{t}^{t+\Delta t} x_n\,dt + O(\Delta t^2) \tag{5.2}$$

$$(5.3) \qquad \epsilon_{\dot{c}} = \epsilon_{\dot{c}_o} - a\,\epsilon_{c_o}\,\Delta t + \int_t^{t+\Delta t} \ddot{x}\,dt + a\int_t^{t+\Delta t} x_n\,dt + O(\Delta t^2)\,.$$

Square Eq. (5.2) and average over the ensemble of inputs $\ddot{x}$, x_n. Then denoting resulting variances and covariances by $\hat{\epsilon}_c$ $\hat{\epsilon}_{\dot{c}}$ $\hat{\epsilon}_{c\dot{c}}$ respectively

$$(5.4) \qquad \hat{\epsilon}_c - \hat{\epsilon}_{c_o} = 2\,\hat{\epsilon}_{c\dot{c}_o}\,\Delta t - 2\,b\,\hat{\epsilon}_{c_o}\,\Delta t + b^2\,\phi\,\Delta t + O(\Delta t^2)$$

where ϕ is the spectral density (assumed constant of x_n). The term ϕ is derived as follows: $\phi(a)$ is the autocorrelation function of x_n.

$$\left\langle \int_t^{t+\Delta t} x_n\,dt \int_t^{t+\Delta t} x_n\,dt \right\rangle = \int_t^{t+\Delta t}\int_t^{t+\Delta t} \langle x_n(u)\,x_n(v)\rangle\,du\,dv$$

$$= \int_o^{\Delta t}\int_o^{\Delta t} \phi(u-v)\,du\,dv$$

$$= \phi \int_o^{\Delta t}\int_o^{\Delta t} \delta(u-v)\,du\,dv$$

$$= \phi\,\Delta t\,.$$

Dividing Eq. (5.4) by Δt and letting $\Delta t \to 0$,

$$\frac{d\hat{\epsilon}_c}{dt} = 2\,\hat{\epsilon}_{c\dot{c}} - 2\,b\hat{\epsilon}_c + b_2\,\phi \,. \tag{5.5}$$

Similarly, assuming the spectral density Θ of $\ddot{x}$ is flat

$$\frac{d\hat{\epsilon}_{\dot{c}}}{dt} = -2\,a\,\hat{\epsilon}_{c\dot{c}} + a^2\,\phi - \Theta \tag{5.6}$$

$$\frac{d\hat{\epsilon}_{c\dot{c}}}{dt} = \hat{\epsilon}_{\dot{c}} - a\,\hat{\epsilon}_c - b\,\hat{\epsilon}_{c\dot{c}} + a\,b\,\phi \,. \tag{5.7}$$

These variance and covariance equations may be used to adjust the gains to get optimum tracking. This will surely result if functions a, b can be found making the right hand members of Eqs. (5.5), (5.6), and (5.7) simultaneously minimum for this will make $\hat{\epsilon}_c$ and $\hat{\epsilon}_{\dot{c}}$ decrease at maximum rate. This simultaneous minimum does occur and at

$$a = \frac{\hat{\epsilon}_{c\dot{c}}}{\phi} \,, \qquad b = -\frac{\hat{\epsilon}_c}{\phi} \,, \tag{5.8}$$

as can be seen by setting the partial derivatives of all three right members with respect to a and b equal to zero. The resulting system has optimum tracking and rate of settling and the variances facilitate evaluating performance of the system.

This optimization of the transient behavior has a by-product-the known steady-state result. For in this case the left hand members of Eqs. (5.5), (5.6), and (5.7) vanish when using Eq. (5.8)

(5.9) $\quad a = \left(\frac{\Theta}{\phi}\right)^{1/2}, \quad b = \sqrt{2}\left(\frac{\Theta}{\phi}\right)^{1/4}. \quad$ (steady-state)

For simplicity the input acceleration spectrum has been assumed flat. There are several ways to deal with non-flat spectra. For example it can be shown that for a general $\Theta(\omega)$ with autocorrelation function $\phi(t)$, Θ in Eq. (5.6) would be replaced by $2\int_0^t W(t, \tau)\, \phi(t - \tau)\, d\tau$. Here $W(t, \tau)$ is the impulsive response of ϵ_c to $\ddot{x}$. Or a flat spectrum could be be filtered by $\Theta(\omega)^{+}[\Theta(\omega) = \Theta(\omega)^{+}\Theta(\omega)^{-}]$ to give a signal of spectrum $\Theta(\omega)$ into the integrator. This last is particularly simple in the Markoff case $\Theta = \frac{\Theta_o}{C^2 + \omega^2}$ where $\Theta^{+} = \frac{1}{C + j\omega}$. The original error tracking loop could be modified as shown in the third figure thus introducing one additional integrator. Proceeding as before six variance - covariance equations result. In general with a system involving n integrators (n^{th} order differential operation) there will be $\frac{n(n+1)}{2}$ variance equations although generally some are trivial.

The transient filtering problem has been discussed on the basis that the noise and signal spectra are known. This leads to a solution of the optimum settling time of a filter and to the best combination of parameters even if the transfer function is not optimal. If the noise and signal spectra are not known then the above technique of computing variances fails and other methods must be used. If sufficient time is available it is possible to measure the spectral densities and use the variance methods but less cumbersome methods are desirable.

VI. ADAPTIVE FILTERS

The discussion which follows will concern adaptive systems, i.e., those which change parameters or adapt as a function of the environment. In general the rate of change of parameters is slow compared to the data rate of the input so that they may be treated as time varying rather than

non-linear systems. If the response of the adapting loop is fast, paper analysis is impossible and simulation techniques are needed. In designing an adaptive system it is necessary to consider the response and stability of the loop as well as the source of intelligence to be employed in adjusting the system.

Adaptive systems may be classified according to several distinct criteria as follows:

1 Object of adaptive loop
   a. Setting of gain or transfer function
   b Adjusting output level or other parameter
   c. Adjusting stability margins of main loop
2. Source of information
   a. Measurements of normal input or output
   b. Injection of tracer signal outside band pass of normal input
   c. Time sharing tracer signal
   d. Amplitude or phase of self excited oscillations
3. Type of system
   a. Open loop, i.e , system adjusted according to measurements on the input
   b. Closed loop, i.e., system measures signal or tracer output.

In addition all adaptive systems may be classified according to standard servo practice as electrical or mechanical, digital or analogue, etc., but such distinctions are not desirable for the present purpose. The class of adaptive servos ranges from standard AGC and AFC loops to servo driven autotransformers for voltage regulation to more sophisticated optimal filtering loops.

Let us now look at examples of the three different adaptive systems listed under criterion 1. These are the zero velocity lag tracking loop mentioned earlier, a similar filter with limited output acceleration, and a system for maintaining a servo loop as tight as possible without instability when the loop gain is slowly varying or not known accurately.

Figure 2 shows a simple servo where only the gain is varied and the loop gain is to be maximized (possibly to minimize the effects of a variable back torque from the load). The unknown gain is assumed to be in the servo. The band-pass filter passes the frequency at which instability is expected and this signal is detected and used to adjust the gain to damp the oscillations. Let λ_o be the gain at which the loop is neutrally stable. Then the rate of buildup or decay of oscillations is proportioned to $(\lambda/\lambda_o - 1)$. The amplitude, z, of the oscillations satisfies the equation

$$\dot{z} = k_1 z(\lambda - \lambda_o)/\lambda_o \doteq k_1 z_o (\lambda - \lambda_o)/\lambda_o \tag{6.1}$$

If the signal is picked off at point A, the control equation is

$$\dot{\lambda} = - g(s) (z - z_o), \tag{6.2}$$

while if it is picked off at point B, we have

$$\dot{\lambda} = g(s) \lambda (z - z_o) \doteq - \lambda_o g(s) (z - z_o) . \tag{6.3}$$

Combining these equations we find that in case A

$$\left[s^2 + \frac{k_1 z_1}{\lambda_o} g(s) \right] \lambda = - k_1 z_1 \lambda_o g(s) \left(\frac{1}{\lambda_o} \right) , \tag{6.4}$$

while for case B

$$(6.5) \qquad \left[s^2 + k_1 z_1 g(s) \right] \lambda = k_1 z_1 g(s) \lambda_o .$$

While both systems can be made stable, and both have $\lambda = \lambda_o$ as the steady state solution, the transient response of the adaptive loop depends on the required gain in case A, contrarily, in case B the transient is fixed and the system has zero velocity lag with respect to variations of λ_o. If λ_o increases linearly with time then $\lambda = \lambda_o$ but $z > z_1$ with $\dot{z} = 0$. Thus a closed loop adaptive system is better here. These statements are subject to modification if there exists dead space or friction in the main or adaptive loop; actually simulation is then required to determine the behavior.

If the main loop shaping network f(s) is properly chosen then a very good servo response is obtainable but input signals at the loop resonance frequency must be avoided. This design is especially useful for a regulator, i.e., when $x_i = 0$ and the servo is designed to counter the back torque.

Figure 3 is a diagram of an adaptive filter in which the RMS output acceleration due to noise is limited. The adaptive loop is very simple and, with the gains as shown, has a response which is independent of input noise spectral density. This can be seen from the fact that the error in RMS acceleration is proportional to $\Delta\lambda/\lambda$ and hence the control equation is

$$(6.6) \qquad \dot{\lambda} \sim \Delta\lambda ,$$

if the filter g(s) is unity. If it is desired to have the adaptive loop time constant proportional to the main loop time constant then the λ must be replaced by λ^2 and the control equation is

$$(6.7) \qquad \dot{\lambda} \sim \lambda \quad \lambda .$$

If the noise is flat or of known spectral shape it is possible to measure its amplitude at the input to the filter and compute the RMS acceleration from the known transfer function of the filter. It is then possible to use an open loop method of adjusting λ but, while this eliminates adaptive loop stability problems it does not have the accuracy in adjusting λ. If the noise does not have the expected spectrum the filter f(s) will not be optimal in shape, however the performance of the system will deviate in the second order if λ is correct, but errors in λ may give first order effects on system performance due to violation of the constraints.

While this loop is very simple we have discussed it because some simulator studies of the effects of non-linearites may be of interest. The non-linearity concerned is that due to a fixed limiter inserted in the loop as shown at the bottom of Figure 3. The analysis was carried out using various fixed values of λ and noise rather than closing the adaptive loop as above. Figure 4 shows the results obtained. The solid upper curve shows σ^2 vs λ for an unlimited system. If the limiter is inserted then the dashed curve is appropriate; the minimum is only a few percent above the optimum at the minimum. The lower half of the figure shows the effect of the limiter on the output acceleration before and after the limit. The result of the simulator study was that the minimum in σ^2 occurs almost exactly at $a_1 = L$, hence the simple adaptive system just described forms a very sophisticated tracking loop.

Figure 5 is a block diagram of the zero velocity lag tracking loop in which we do not know the signal or noise spectra although we assume the noise to be nearly flat. In order to adjust band-pass we may use the result, Eq. (3.13), so that the error signal spectrum is proportional to the noise spectrum when the loop is optimal. While the transfer of the loop is not correct if the noise is not flat or if the signal is not that assumed, it is still true that adjusting the loop band-pass so that the error signal spectrum is flat is nearly optimum.

The method of measurement is to use two filters $f_1(s)$, a low pass filter covering the band-pass of the main loop, and $f_2(s)$ covering an equal band-pass just above the main loop and take the ratio of the outputs. The optimum shape of such filters has not been determined but simulator runs show suitable performance for simple filters. $f_2(s)$ should have a finite band-pass because the actual high frequency noise is unimportant; only the noise in the vicinity of the main loop band-pass is important.

In the loop as shown the obvious scale factors have been inserted to make the response frequency - relative to the main loop - independent of the value of the spectra. The filter $f_3(s)$ determines the band-pass of the adaptive loop and, in order to have minimum RMS errors in λ, $f_3(s)$ must be adjusted so that the lags in following changes in the spectra are balanced by the fluctuations in the noise out of the detectors. From this criterion we can determine the adaptive loop band-pass as $\omega_A \sim \sqrt{\frac{\lambda}{T}}$, where $1/T \sim \dot{\phi}/\phi$ is the effective time constant relating to the change in input spectra. However, if step changes in the input signals are contemplated then the adaptive loop should be as tight as stability dictates and the gain settings in the figure are correct. The ratio of the filter outputs minus one is proportional to $\delta\lambda/\lambda$ so that

$$\dot{\lambda} \sim \lambda \quad \delta\lambda \tag{6.8}$$

and the band-pass in the adaptive loop is proportional to that in the main loop.

In all of the adaptive systems considered it is easy to specify the gain changes to keep the loop dynamically similar for different inputs, but it is harder to specify the exact band-pass or the shape of the filters in the adaptive loop. It is always possible to make a linear stability analysis, if the inputs are fixed, and the noise out of the squaring circuits can be computed for Markovian noise but no general theory of optimal design exists.

There are many ways of instrumenting adaptive servos which give adequate performance and the effects of non-linearites and complexity must be considered carefully if a satisfactory design is to be obtained. For example the use of smoothed absolute value instead of RMS leads to only a few per cent more noise in the adaptive loop. As another example the division in the last example may be replaced by a subtraction if the dynamic range is not too great. At low input signals the loop is then sluggish but the tracking error is small due to the small input.

Underlying the design of adaptive servos is the assumption of relatively slow, or only occasional, changes

in environment. If rapid changes occur a different main loop, possibly non-linear, is required. Adaptive loops may be called quasi-linear but because they are non-linear no general method of analysis has emerged to determine optimal performance as a standard of comparison with specific loops, or to check instrumentation approximations.

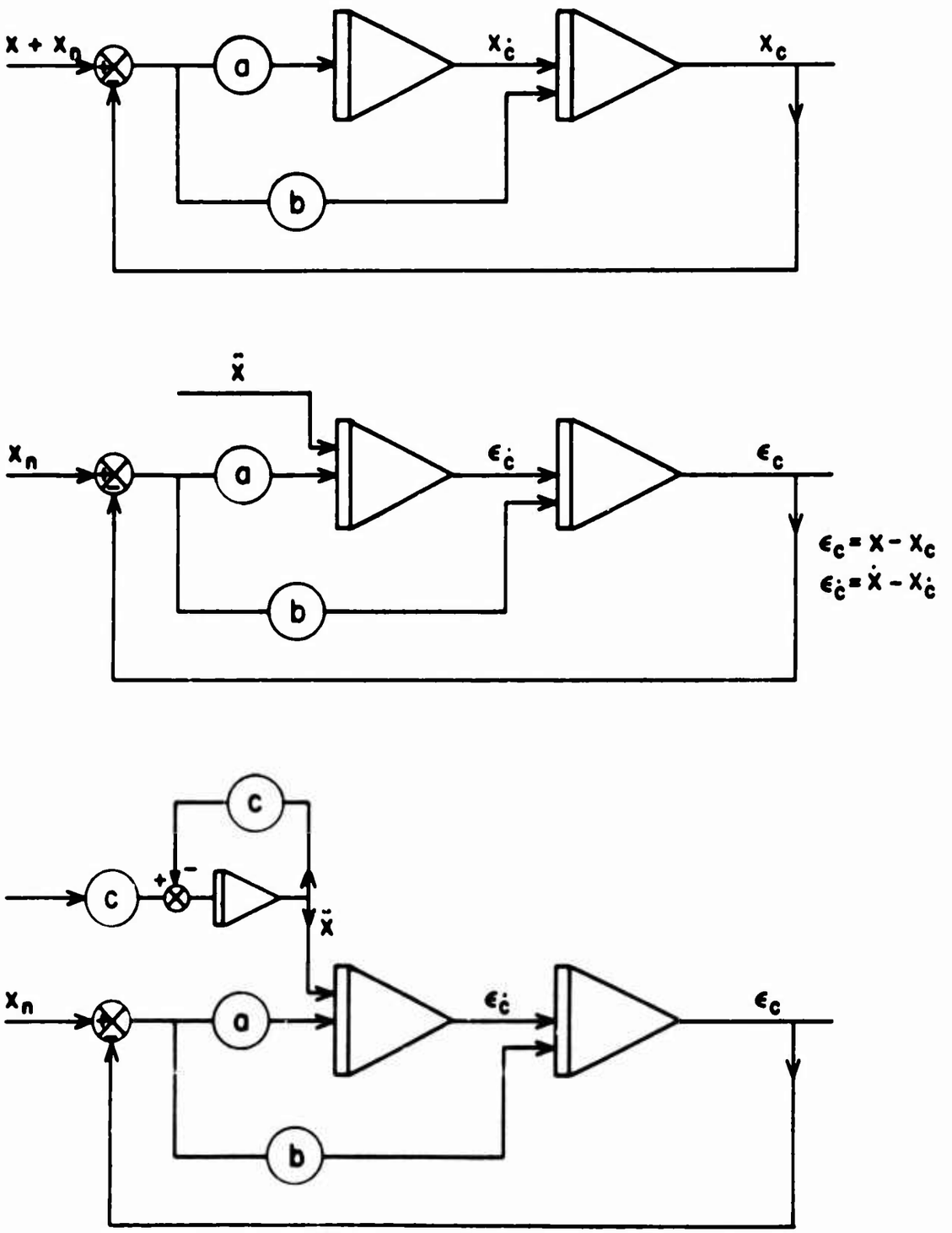


Fig. 1 ZERO VELOCITY LAG FEEDBACK SYSTEM

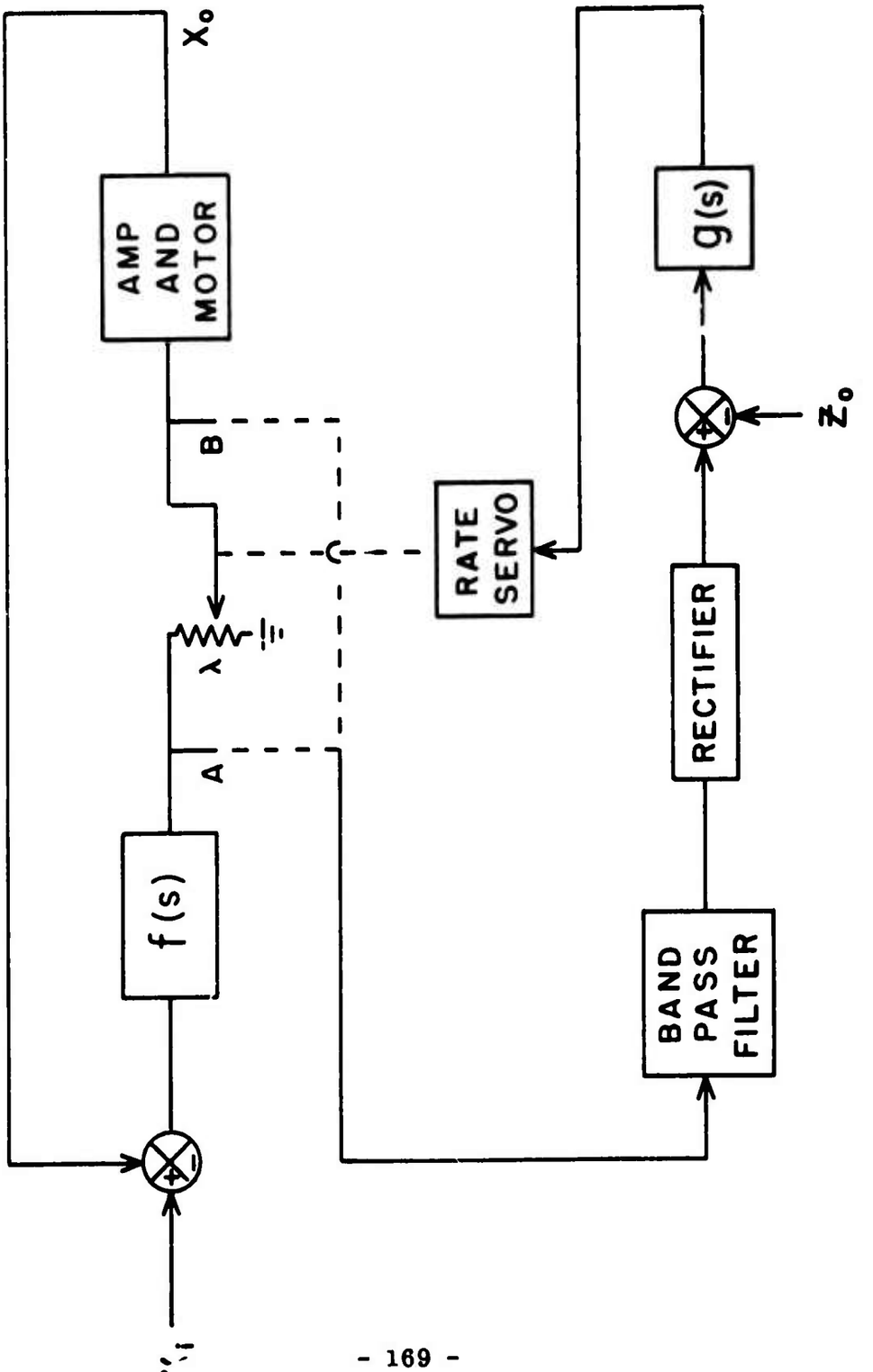


Fig. 2 MAXIMIZED LOOP GAIN SERVO

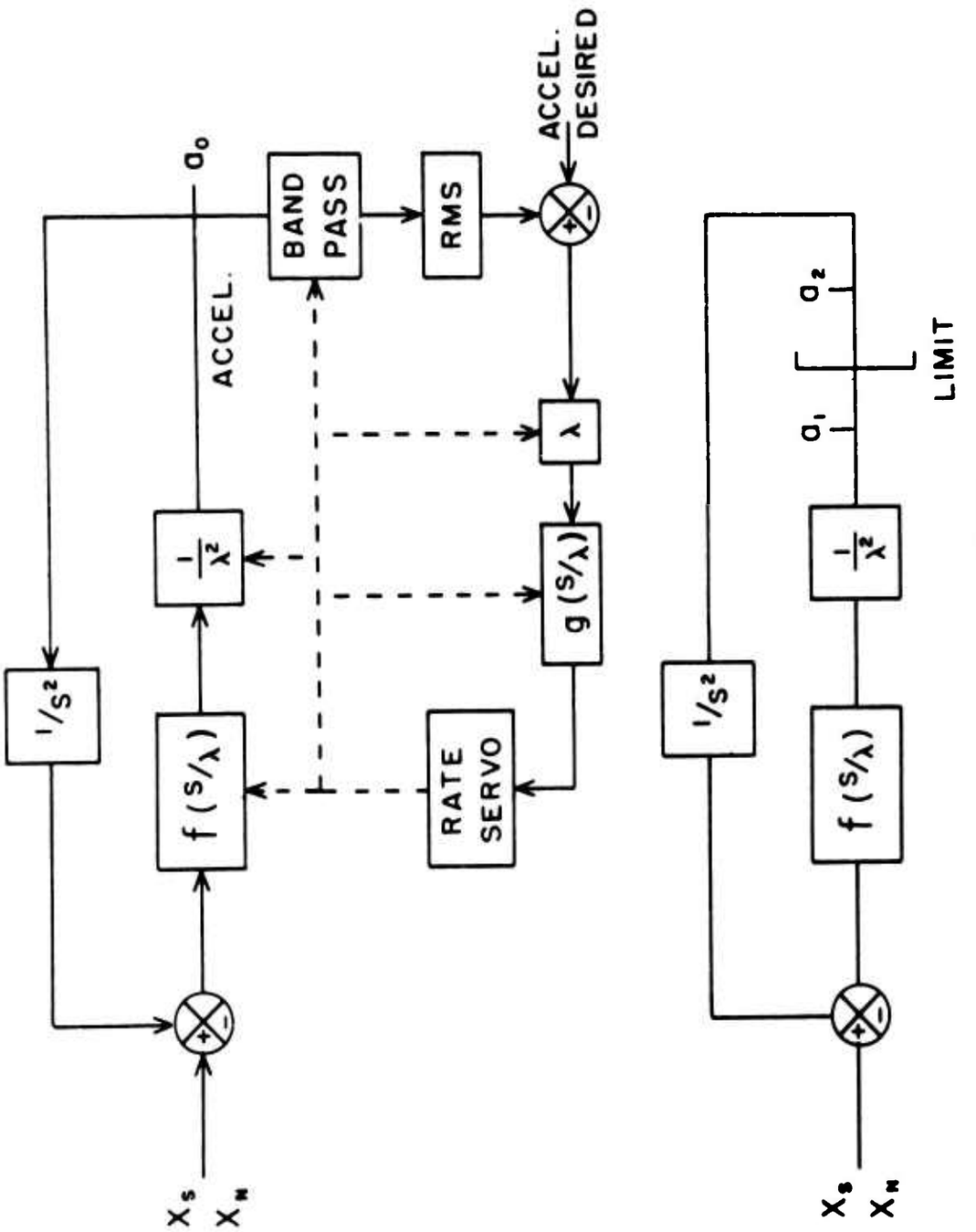


Fig. 3 AN ADAPTIVE FILTER

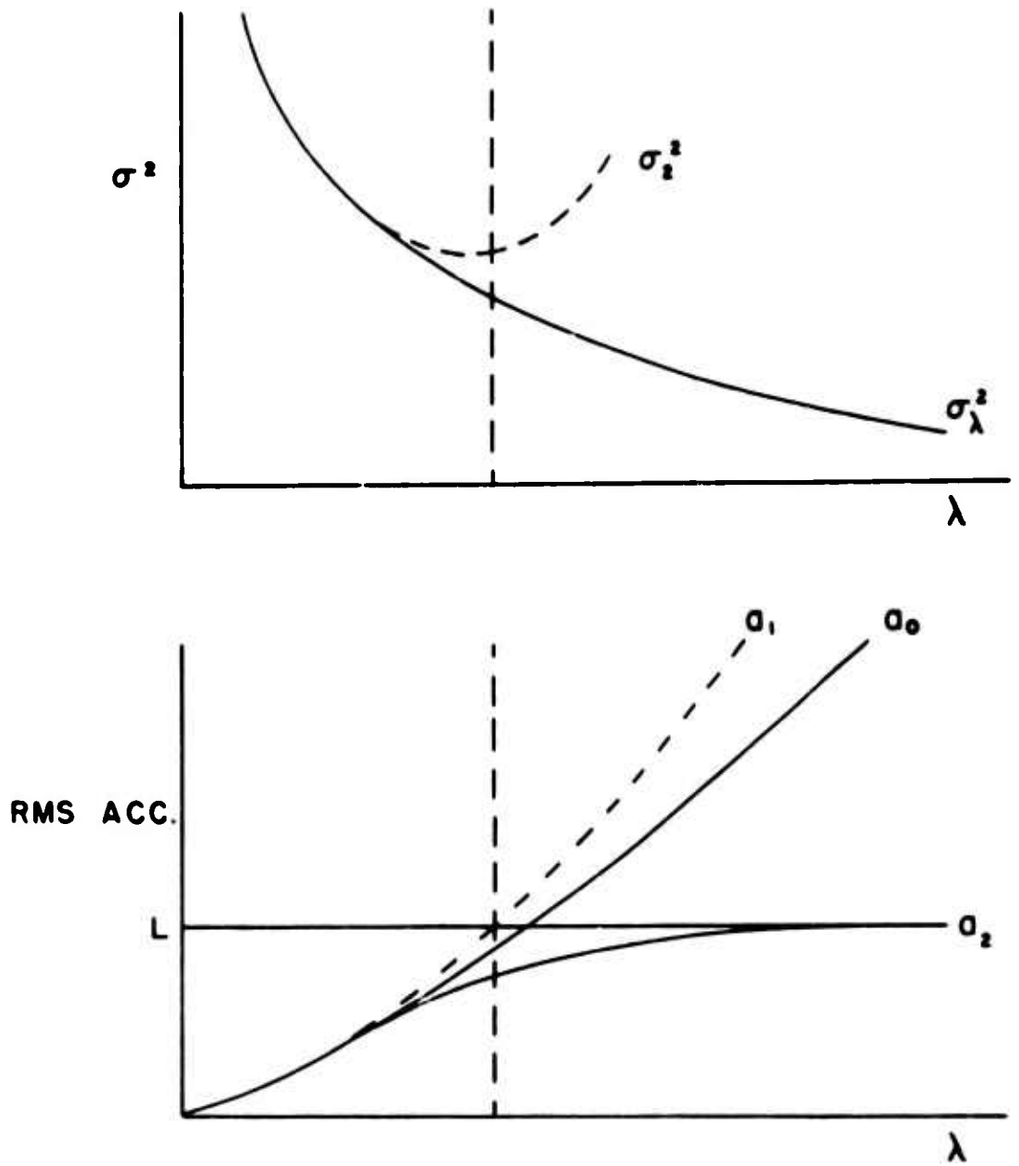


Fig. 4 RESULTS OF FIG. 3

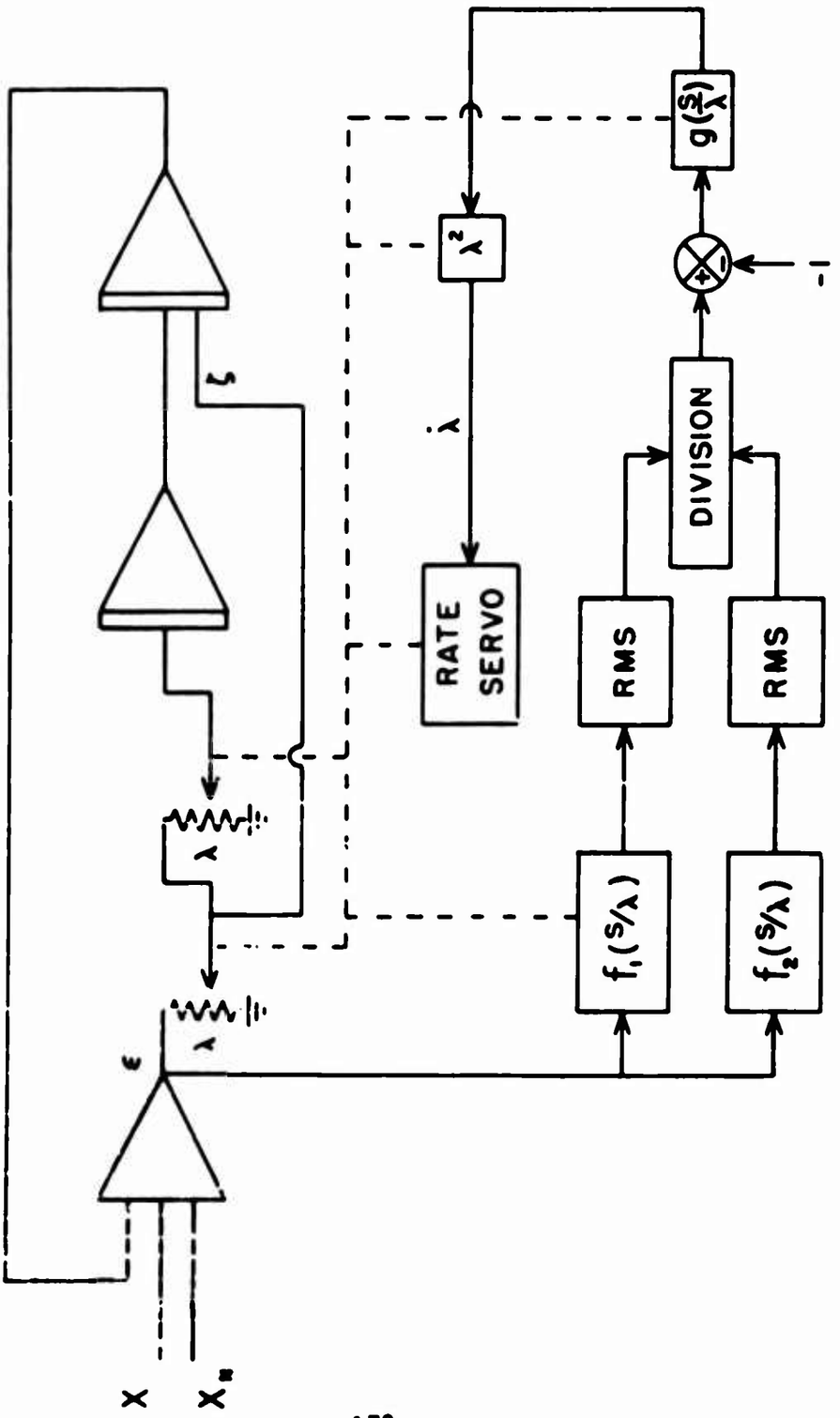


Fig. 5 ZERO VELOCITY LAG TRACKING LOOP

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

OPTIMAL FILTERING IN MISSILE GUIDANCE

by

A. George Carlton

I. INTRODUCTION

The guidance and control system of a guided missile consists of three basic elements: a guidance intelligence system, a guidance computer, and an autopilot. The guidance intelligence system measures in a suitable reference frame work the position of the missile relative to the target, the autopilot causes the missile to execute maneuvers as commanded, and the guidance computer commands maneuvers on the basis of inputs received from the guidance intelligence system.[1]

The guidance computer cannot simply command an acceleration proportional to the measured error in missile position, as the resulting change in missile position would affect the measured error so as to result in harmonic oscillation of the missile. This can be corrected by adding to the command a term proportional to the error derivative, damping the oscillation. With proper choice of the constants, such a guidance computer may cause the missile to respond very faithfully to the inputs from guidance intelligence. This is not typically sufficient to result in the smallest possible misses, however, because the guidance intelligence is not perfect. In addition to genuine information as to error of missile from target, called the signal, the guidance intelligence gives spurious information called noise. There are many sources of noise, including internal noise in the sensing device, atmospheric effects, and scintillations of the target. The missile should respond to the signal but ignore the noise. The problem of optimal filtering in missile guidance is the problem of achieving the best compromise between these conflicting requirements.

It is sometimes thought that the noise effects are of minor importance. Consider a very simplified situation in which the only elements are acceleration capability of the target and the level of noise which is "white", i.e., uniformly spread over all frequencies. Simple dimensional analysis shows that the minimum mean square miss is proportional to the 4/5 power of the noise spectral density, and a little more effort shows that 80 per cent of the minimum mean square miss is due to following the noise.

---

1) Presented at Indianapolis Meeting, AAAS, 28 December 1957.

The very similar problem in directing antiaircraft gunfire was one application considered by Wiener in "Extrapolation, Interpolation, and Smoothing of Stationary Time Series" Wiener considered signal and noise ensembles which were stationary stochastic processes and determined the realizable linear filter which minimized the mean square miss. If the signal and noise are uncorrelated, the characteristics of interest are the power spectral densities $S(\omega)$ and $N(\omega)$ representing the resolution of the mean square signal and the mean square noise into a continuum of frequency components. Thus $E\,[\text{signal}]^2 = \int S(\omega)\, d\omega$. The linear filter is defined by the transfer function $F(\omega)$ specifying the scale factor and phase shift applied to each frequency. The mean square error σ^2 is given by

$$(1) \qquad \sigma^2 = \int \left\{ \left| 1 - F \right|^2 S + \left| F \right|^2 N \right\} d\omega \quad .$$

since $|1 - F|$ is the erroneous factor applied to signal components and $|F|$ the erroneous factor on noise components. To be <u>realizable</u>, the filter must operate only on inputs from the past, not on future inputs. The realizable F minimizing σ^2 was found by Wiener to be

$$(2) \qquad F_{S,N}(\omega) = \frac{1}{2\pi\,[S(\omega) + N(\omega)]^+} \int_0^\infty e^{-i\omega t} dt \int_{-\infty}^{\infty} \frac{S(u)\, e^{iut}\, du}{[S(u) + N(u)]^-} \quad .$$

In this expression, $S(\omega) + N(\omega)$ is factored into $(S + N)^+$ $(S + N)^-$, $(S + N)^+$ being analytic and without poles or zeroes in the lower half-plane, with $(S + N)^-$ the complex conjugate of $(S + N)^+$.

These results have been applied directly to missile guidance problems, but have some serious shortcomings. The most obvious defect is that the spectral density of the signal

is assumed fixed and known Since an enemy interested in increasing our error may select the signal characteristics, it is reasonable to assume varied techniques chosen to impair the performance of our system

II FILTERING PROBLEM

The problem of filtering is essentially one of statistical decision. Statistical decision theory in this problem, whether we are fighting a conscious enemy or implacable nature, calls for selecting the filter for which the σ^2 maximized over possible signal spectra is a minimum.

In view of the complexity of Eq. (2), the determination of this minimax filter appears difficult The minimax filter can be determined without too much trouble, however, by the use of indirect approaches. The first step in solving the problem is to investigate the maximin $S(\omega)$, that is, the $S(\omega)$ which maximizes the minimum achievable σ^2.

The possible signal spectra will be limited by linear restrictions, limits on the power of linear functions of the signal. An airplane, for example, will have some bounds on its position, velocity, and acceleration, due to limits on its course, propulsion, and structure. If the mean square acceleration cannot exceed a^2, then $\int S(\omega)\ \omega^4\ d\omega \leq a^2$. In general, there may be n such linear restrictions of the form

$$(3) \qquad k_j \int S(\omega)\ \Theta_j(\omega)\ d\omega = 1 \ ,$$

with $k_j > 0$, $S(\omega) \geq 0$, $\Theta_j(\omega) \geq 0$.

For example, with limited mean square velocity and acceleration, we might have $\Theta_1 = \omega^2$, $\Theta_2 = \omega^4$.

The maximin $S(\omega)$ is that $S(\omega)$ subject to these constraints which maximizes

$$(4) \qquad \min_F \sigma^2_{S,N} = \int \left[\left| F_{S,N} \right|^2 N + \left| 1 - F_{S,N} \right|^2 S \right] d\omega .$$

Introducing the constraints by Lagrange's method, we wish to maximize

$$(5) \qquad I_{S,N} = \min_{F} \sigma^2_{S,N} + \sum_j \lambda_j k_j \int S(\omega)\, \theta_j(\omega)\, d\omega =$$

$$= \int \left\{ \left| F_{S,N} \right|^2 N + \left[\left| 1 - F_{S,N} \right|^2 - \sum_j \lambda_j k_j \theta_j \right] S \right\} d\omega \ .$$

For any increment Δ on S_j the resulting increment on I is

$$(6) \qquad \delta I = \int \left\{ \left| F_{S+\Delta,N} \right|^2 N + \left| 1 - F_{S+\Delta,N} \right|^2 (S + \Delta) \right\} d\omega$$

$$- \int \left\{ \left| F_{S,N} \right|^2 N + \left| 1 - F_{S,N} \right|^2 (S + \Delta) \right\} d\omega$$

$$+ \int \Delta \left\{ \left| 1 - F_{S,N} \right|^2 - \sum_j \lambda_j k_j \theta_j \right\} d\omega \ .$$

For a maximin $S(\omega)$ it is required that δI be nonpositive for all permissible Δ.

The first integral is the minimum σ^2 given spectra $S + \Delta$ and N, the second integral and σ^2 given the same spectra and nonoptimal F, so the difference of the two integrals is nonpositive. Thus δI is not greater than the third integral. This integral will be nonpositive if the coefficient of Δ is zero when a Δ of either sign is permissable and nonpositive when any permissible Δ is non-negative. Since $S(\omega) \geq 0$, permissible Δ are of either sign for $S(\omega) > 0$ but are non-negative when $S(\omega) = 0$.

Thus the maximin $S(\omega)$ is $S_o(\omega)$ such that

$$(7) \qquad \left| 1 - F_{S_o,N} \right|^2 = \sum_j \lambda_j k_j \theta_j \qquad \omega : S_o(\omega) > 0$$

$$\leq \sum_j \lambda_j k_j \theta_j \qquad \omega : S_o(\omega) = 0,$$

with the λ_j chosen to satisfy the given constraints. The result just obtained is unsatisfactory in that the constraints were applied strictly, rather than as inequalities By similar manipulation one can show that when the unique constraints are replaced by inequality constraints, $k_j \geq 1$, the result is the same as above, together with the condition

$$(8) \qquad \lambda_j (k_j - 1) = 0,$$

implying that each constraint is either redundant or applied strictly.

To derive the minimax filter, we show that the game is determined since

$$\min_F \max_S \sigma^2_{S,N,F} \leq \max_S \sigma^2_{S,N,F_{S_o,N}} =$$

$$(9) \qquad = \max_S \int \left\{ N \left| F_{S_o,N} \right|^2 + S \left| 1 - F_{S_o,N} \right|^2 \right. d\omega =$$

$$= \int N \left| F_{S_o,N} \right|^2 d\omega + \sum_j \lambda_j = \max_S \min_F \sigma^2_{S,N,F} .$$

Thus the minimax σ^2 is equal to or less than the minimin σ^2. Since the fundamental theorem of game theory states that the minimax is equal to or greater than the maximin, the game is determined, with $F_{S_o,N}$ the minimax filter.

The second indirect approach is to deal with the function $\left|1 - F_{S,N}\right|^2$ which is much more tractable than $F_{S,N}$ or the formally equivalent $1 - F_{S,N}$. Applying variational procedures to $\sigma^2 = \int \left\{ \left|F\right|^2 N + \left|1 - F\right|^2 S \right\} d\omega$, one can obtain the result derived by Wiener for the optimal realizable $F_{S,N}$ and also the useful result:

$$(10) \qquad \left|1 - F_{S,N}\right|^2 (S + N) = \mathrm{Re}N\bar{F} + \text{Real complement to } \mathrm{Im}\, N\bar{F} \ ,$$

where the real complement is the real function which must supplement the imaginary function to yield a realizable transfer function. For simple forms of N, Eq. (10) can be solved rather easily, for example

$$\left|1 - F_{S,N}\right|^2 (S + N), = N \qquad N(\omega) \text{ constant}$$

$$N\left[1 - \bar{F}(ia)\right], \quad N(\omega) = \frac{ca^2}{a^2 + \omega^2}$$

(11)

$$N + K, \qquad N(\omega) = C_o + C_2\omega^2 + C_4\omega^4$$

$$\left(K \text{ determined by } \int \log \left|1 - F_{S,N}\right|^2 d\omega = 0\right).$$

Divergent spectra play a useful role in some problems and the general results on optimal filters are still valid.

By using the results Eq. (7) and (11), it is possible to determine $S_o(\omega)$ and $\left|1 - F_{S_o,N}\right|^2$ for the minimax filter, and one can then determine $F_{S_o,N}$ by factorization in special cases or in general by applying Wiener's result to the determined $S_o(\omega)$.

To illustrate the method, consider the case of white noise of spectral density N and mean square target acceleration equal to or less than a^2. Our constraint is then

$$\frac{1}{a^2} \int s(\omega)\ \omega^4\ d\omega = 1.$$

From Eq. (7),

$$\left|1 - F\right|^2 = \lambda \frac{\omega^4}{a^2} \qquad \omega: S(\omega) > 0$$

$$\leq \lambda \frac{\omega^4}{a^2} \qquad \omega : S(\omega) = 0 .$$

From Eq. (11),

$$\left|1 - F\right|^2 = \frac{N}{S + N} .$$

Thus, setting $a^2/\lambda = \omega_o^4$

$$\frac{N}{S + N} = \begin{cases} \dfrac{\omega^4}{\omega_o^4} & \omega : S(\omega) > 0 \\ 1 & \omega : S(\omega) = 0 . \end{cases}$$

The solution for S is

$$S = N\left(\frac{\omega_o^4}{\omega^4} - 1\right) , \qquad \omega^2 < \omega_o^2$$

$$0 , \qquad \omega^2 \geq \omega_o^2 ,$$

ω_o must satisfy the relation $\frac{1}{a^2} \int S(\omega)\ \omega^4\ d\omega = 1$, or

$$-\int_{-\omega_o}^{\omega_o} N\ (\omega_o^4 - \omega^4)\ d\omega = a^2, \text{ from which } \omega_o = \left(\frac{5}{8}\frac{a}{N}\right)^{2\ 1/5}$$

It is clear that $\left|1 - F\right|^2 = \omega^4/\omega_o^4, \ \omega^2 < \omega_o^2$

$1 \quad , \ \omega^2 > \omega_o^2$.

Some numerical calculation is required to obtain F itself.

The minimax mean square miss is obtained from the relation, valid for white noise,

$\sigma^2 = -\int N \log \left|1 - F_{S,N}\right|^2 d\omega$, from which we find σ^2 $= 8 N \omega_o$. The portion of σ^2 due to failure to follow the signal is $\int S \left|1 - F\right|^2 d\omega = 8 N \omega_o/5$.

From the fact that $\left|1 - F\right|^2$ is unity for all large ω, it follows that $F(\omega)$ is proportional to $1/\omega$ for large frequencies. From the practical standpoint of guided missile design, this is most unfortunate, as the mean square acceleration of the missile resulting from noise is

$$\int N \left|F\right|^2 \omega^4 \, d\omega$$

which clearly does not converge for N constant and F of order $1/\omega$.

It is therefore necessary to modify our problem so as to select a minimax filter among those for which $\int N \left|F\right|^2 \omega^4 d\omega$ is limited to a specified value. To solve this problem we observe that the previous determination of the maximin S is valid regardless of the class over which F is minimized, so we need merely find the optimal F among those filters for which $\int N \left|F\right|^2 \omega^4 d\omega$ does not exceed a specified value. The quantity to be minimized becomes

$$\sigma^2_{S,N,F} + \lambda \int N \left|F\right|^2 \omega^4 \, d\omega =$$

$$= \int \left\{ S \left|1 - F\right|^2 + N \left|F\right|^2 (1 + \lambda \omega^4) \right\} d\omega,$$

with λ chosen to satisfy the mean square missile acceleration requirement. This is exactly equivalent to the former filter optimization problem with an effective noise spectral density of $N(1 + \lambda \omega^4)$. It is to solve problems of this sort that one considers non-convergent noise spectra, such as $N = N_o + C_2 \omega^2 + C_4 \omega^4$. Calculations of minimax filter for such cases have been carried out. The optimal filters and the minimax mean square miss depend on the ratio of possible mean square target acceleration to allowable mean square missile acceleration due to noise. This ratio should be noticeably larger than unity to avoid significant decrease in accuracy.

A completely different approach to the problem is hinted at by some of the results obtained for optimal $|1 - F|^2 (S + N)$. This quantity is the spectral density of the difference between the input and the output of the filter, a difference which is the error signal in a servo type filter. The spectral density of this difference is equal to the noise spectral density if the noise is white and the filter is optimum. Useful but less simple relations can be derived for other types of noise. Thus we have the possibility of checking the correctness of the filtering process by investigating the spectrum of this difference, and adjusting the filter accordingly. An advantage of this type of filtering is that it can take advantage of a situation in which the opponent's strategy is poor. With sufficiently slow adaptation, such a filter will closely approximate the minimax filter in the most difficult situation, and can be considered a sub-minimax filter, one of the few practical examples of the existence of sub-minimax solutions to game theory problems.

Adaptive filters based on these considerations have been devised and have demonstrated excellent performance, although there is at present no adequate theory in this field.

THE JOHNS HOPKINS UNIVERSITY
APPLIED PHYSICS LABORATORY
SILVER SPRING MARYLAND

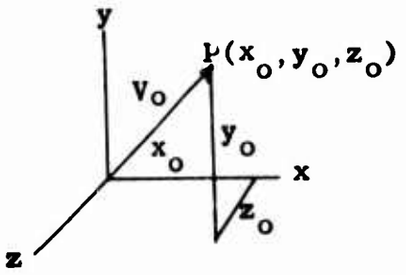
THREE DIMENSIONAL COORDINATE SYSTEMS AND MISSILE DYNAMICS

by

J. E. Hanson

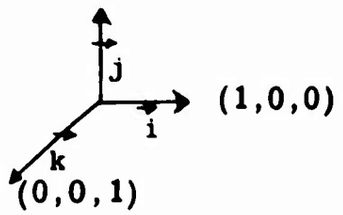
I. SUMMARY OF USEFUL FORMULAS IN VECTOR ANALYSIS

Vector Representation---Let x, y, z be a right-handed three-dimensional rectangular coordinate system. A vector $\vec{V}_o$ can be represented as an ordered triple of numbers corresponding to the coordinates of the point P. Two concepts characterize a vector: magnitude and direction. A vector can be thought of as a directed line segment, i.e., with a head on one end, and a tail on the other. Translations of the vector do not alter the vector. Thus, if the tail is located at (x_1, y_1, z_1), the head at (x_2, y_2, z_2), the vector is represented by the ordered triple $(x_2-x_1, y_2-y_1, z_2-z_1)$, which does not depend on where the tail is located.



A more useful representation of a vector is based on the concept of base vectors. Let $\vec{i}$, $\vec{j}$, $\vec{k}$ denote the vectors whose ordered triple representations are (1, 0, 0), (0, 1, 0), (0, 0, 1), respectively. Such a triad is called a right-handed orthogonal triad of unit vectors, and forms a coordinate system. In terms of these unit vectors, the vector $\vec{V}_o$ above is expressed (0, 1, 0) as

$$(1.1) \quad \vec{V}_o = x_o \vec{i} + y_o \vec{j} + z_o \vec{k} .$$



Length of a Vector---The length of the vector $\vec{V}_o$, denoted by $|\vec{V}_o|$, is given by

$$(1.2) \quad |\vec{V}_o| = \sqrt{x_o^2 + y_o^2 + z_o^2} .$$

Addition of Vectors---If $\vec{V}_o = x_o\vec{i} + y_o\vec{j} + z_o\vec{k}$, $V_1 = x_1\vec{i} + y_1\vec{j} + z_1\vec{k}$, then $\vec{V}_o + \vec{V}_1$ is defined to be a vector given by

$$(1.3)\quad \vec{V}_o + \vec{V}_1 = (x_o + x_1)\vec{i} + (y_o + y_1)\vec{j} + (z_o + z_1)\vec{k}\ .$$

This definition is in agreement with the parallelogram law found in most elementary physics texts.

Multiplication of a Vector by a Scalar (Number)--- $\alpha\vec{V}_o$ is defined to be the vector given by

$$(1.4)\quad \alpha\vec{V}_o = \alpha x_o\vec{i} + \alpha y_o\vec{j} + \alpha z_o\vec{k}\ .$$

Exercises: Prove

(a) $\vec{V}_o + \vec{V}_1 = \vec{V}_1 + \vec{V}_o$

(b) $\vec{V}_o + (\vec{V}_1 + \vec{V}_2) = (\vec{V}_o + \vec{V}_1) + \vec{V}_2$

(c) $\alpha(\vec{V}_o + \vec{V}_1) = \alpha\vec{V}_o + \alpha\vec{V}_1$

Dot Product (or Scalar Product) of Two Vectors--- The dot product of two vectors is a scalar.

It is denoted by $\vec{V}_o \cdot \vec{V}_1$, and is defined by

$$(1.5)\quad \vec{V}_o \cdot \vec{V}_1 = |\vec{V}_o|\ |\vec{V}_1| \cos\phi\ ,$$ where ϕ is the angle between the two vectors.

It can be shown that

$$(1.6)\quad \vec{V}_o \cdot \vec{V}_1 = x_o x_1 + y_o y_1 + z_o z_1\ .$$

Cross Product (or Vector Product) of Two Vectors---The cross product of two vectors is a vector.

It is denoted by $V_o \times V_1$, and is defined by

$$(1.7)\quad \vec{V}_o \times \vec{V}_1 = |V_o| \; |V_1| \sin \phi \; \vec{n},$$

where ϕ is as above, $\vec{n}$ is a unit vector normal to the plane formed by $\vec{V}_o$ and $\vec{V}_1$, its direction determined by the right-hand rule. It can be shown that

$$(1.8)\quad \vec{V}_o \times \vec{V}_1 = (y_o z_1 - z_o y_1)\vec{i} + (z_o x_1 - z_1 x_o)\vec{y} + (x_o y_1 - y_o x_1)\vec{k}$$

or, in the form of a determinant,

$$(1.9)\quad \vec{V}_o \times \vec{V}_1 = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ x_o & y_o & z_o \\ x_1 & y_1 & z_1 \end{vmatrix} .$$

Components---The component of a vector in a direction is a vector formed by projecting the vector onto a line pointing in that direction. Thus, the component of $\vec{V}_o$ in the $\vec{i}$ direction is $x_o \vec{i}$.

Exercises: Prove

(d) $\vec{V}_o \cdot \vec{V}_1 = \vec{V}_1 \cdot \vec{V}_o$.

(e) $\vec{V}_o \cdot (\vec{V}_1 + \vec{V}_2) = \vec{V}_o \cdot \vec{V}_1 + \vec{V}_o \cdot \vec{V}_2$.

(f) $\alpha(\vec{V}_o \cdot \vec{V}_1) = (\alpha\vec{V}_o) \cdot \vec{V}_1 = \vec{V}_o \cdot (\alpha\vec{V}_1)$.

(g) $\vec{V}_o \times \vec{V}_1 = -(\vec{V}_1 \times \vec{V}_o)$.

(h) $\alpha(\vec{V}_o \times \vec{V}_1) = (\alpha\vec{V}_o) \times \vec{V}_1 = \vec{V}_o \times (\alpha\vec{V}_1)$.

(i) $\vec{V}_o \times (\vec{V}_1 \times \vec{V}_2) = (\vec{V}_o \cdot \vec{V}_2)\vec{V}_1 - (\vec{V}_o \cdot \vec{V}_1)\vec{V}_2$.

(i') $\vec{V}_o \times (\vec{V}_1 + \vec{V}_2) = (\vec{V}_o \times \vec{V}_1) + (\vec{V}_o \times \vec{V}_2)$

(j) $\vec{V}_o \cdot (\vec{V}_1 \times \vec{V}_2) = (\vec{V}_o \times \vec{V}_1) \cdot \vec{V}_2 =$

$$\begin{vmatrix} x_o & y_o & z_o \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}$$

(This expression is sometimes called the box product of three vectors.)

(k) The component of $\vec{V}_o$ in the $\vec{V}_1$ direction (if $|\vec{V}_1| \neq 0$) is given by

$$\left(\frac{\vec{V}_o \cdot \vec{V}_1}{|\vec{V}_1|} \right) \frac{\vec{V}_1}{|\vec{V}_1|} .$$

(l) $\vec{V}_o \cdot \vec{i} = x_o$, $\vec{V}_o \cdot \vec{j} = y_o$, $\vec{V}_o \cdot \vec{k} = z_o$, so that one can write $\vec{V}_o = (\vec{V}_o \cdot \vec{i})\vec{i} + (\vec{V}_o \cdot \vec{j})\vec{j} + (\vec{V}_o \cdot \vec{k})\vec{k}$.

(m) Let $\vec{V}_1$, $\vec{V}_2$ be non-collinear, and different from zero.

Let $\vec{V}_3$ be the vector formed by projecting $\vec{V}_o$ into the plane formed by $\vec{V}_1$ and $\vec{V}_2$. Then $\vec{V}_3 = \vec{n} \times (\vec{V}_o \times \vec{n})$, where $\vec{n} = \dfrac{\vec{V}_1 \times \vec{V}_2}{|\vec{V}_1 \times \vec{V}_2|}$.

The following are in general impossible or incorrect statements. Why?

(n) $\vec{V}_o \times (\vec{V}_1 \times \vec{V}_2) = (\vec{V}_o \times \vec{V}_1) \times \vec{V}_2$.

(o) $\vec{V}_o \cdot (\vec{V}_1 \times \vec{V}_2) = (\vec{V}_o \cdot \vec{V}_1) \times \vec{V}_2$.

(p) $\vec{V}_o \cdot (\vec{V}_1 \cdot \vec{V}_2) = (\vec{V}_o \cdot \vec{V}_1) \cdot \vec{V}_2$.

II. ROTATED COORDINATE SYSTEMS

Let $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ denote a right-handed orthogonal triad of unit vectors distinct from $\vec{i}$, $\vec{j}$, $\vec{k}$. Expressing each $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ as the sum of its components along $\vec{i}$, $\vec{j}$, $\vec{k}$, we can write

$$(2.1)\quad \begin{aligned} \vec{i}_1 &= \alpha_{11}\,\vec{i} + \alpha_{12}\,\vec{j} + \alpha_{13}\,\vec{k} \\ \vec{j}_1 &= \alpha_{12}\,\vec{i} + \alpha_{22}\,\vec{j} + \alpha_{23}\,\vec{k} \\ \vec{k}_1 &= \alpha_{31}\,\vec{i} + \alpha_{32}\,\vec{j} + \alpha_{33}\,\vec{k} \end{aligned} .$$

Since $|\vec{i}_1| = 1$, $|\vec{j}_1| = 1$, $\vec{i}_1 \cdot \vec{j}_1 = 0$, $\vec{i}_1 \times \vec{j}_1 = \vec{k}_1$, we have

$$(2.2)\quad \alpha_{11}^2 + \alpha_{12}^2 + \alpha_{13}^2 = 1,$$

$$(2.3)\quad \alpha_{21}^2 + \alpha_{22}^2 + \alpha_{23}^2 = 1,$$

$$(2.4)\quad \alpha_{11}\,\alpha_{21} + \alpha_{12}\,\alpha_{22} + \alpha_{13}\,\alpha_{23} = 0,$$

$$(2.5)\quad \alpha_{31} = \alpha_{12}\,\alpha_{23} - \alpha_{22}\,\alpha_{13},$$

$$(2.6)\quad \alpha_{32} = \alpha_{13}\,\alpha_{21} - \alpha_{11}\,\alpha_{23},$$

$$(2.7)\quad \alpha_{33} = \alpha_{11}\,\alpha_{22} - \alpha_{21}\,\alpha_{12}\,.$$

Conversely, it can be shown that any set of α's satisfying Eqs. (2.2) - (2.7) will produce a right-handed orthogonal triad $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ when substituted into Eq. (2.1).

If the α's are as above, and one solves Eq. (2.1) for $\vec{i}$, $\vec{j}$, $\vec{k}$ in terms of $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$, the solution, as can be shown, is expressible in the following simple form:

$$(2.8)\quad \begin{aligned} \vec{i} &= \alpha_{11}\,\vec{i}_1 + \alpha_{21}\,\vec{j}_1 + \alpha_{31}\,\vec{k}_1 \\ \vec{j} &= \alpha_{12}\,\vec{i}_1 + \alpha_{22}\,\vec{j}_1 + \alpha_{32}\,\vec{k}_1 \\ \vec{k} &= \alpha_{13}\,\vec{i}_1 + \alpha_{23}\,\vec{j}_1 + \alpha_{33}\,\vec{k}_1 \end{aligned} .$$

Equations (2.1) and (2.8) make the problem of expressing in a second coordinate system a vector known in one coordinate system quite simple:

Thus

$$
\begin{aligned}
(2.9)\quad \vec{V}_o &= x_o \vec{i} + y_o \vec{j} + z_o \vec{k} = x_o (\alpha_{11} \vec{i}_1 + \alpha_{21} \vec{j}_1 + \alpha_{31} \vec{k}_1) \\
&\quad + y_o (\alpha_{12} \vec{i}_1 + \alpha_{22} \vec{j}_1 + \alpha_{32} \vec{k}_1) \\
&\quad + z_o (\alpha_{13} \vec{i}_1 + \alpha_{23} \vec{j}_1 + \alpha_{33} \vec{k}_1) \\
&= (x_o \alpha_{11} + y_o \alpha_{12} + z_o \alpha_{13}) \vec{i}_1 \\
&\quad + (x_o \alpha_{21} + y_o \alpha_{22} + z_o \alpha_{23}) \vec{j}_1 \\
&\quad + (x_o \alpha_{31} + y_o \alpha_{32} + z_o \alpha_{33}) \vec{k}_1 .
\end{aligned}
$$

III. DERIVATIVES OF VECTORS AND MOVING COORDINATE SYSTEMS

Suppose the vector V is a function of time. Since a vector can be described by its components, it is clear that a situation could arise where the components are constants in one coordinate system, yet varying in another coordinate system moving with respect to the first. To talk about the derivative of a vector, we must then first specify the coordinate system. In Newtonian Mechanics, most physical laws expressible in vectorial form achieve their simplest expressions when the coordinate system is fixed relative to the fixed stars (or moving with a uniform velocity with respect to them). (In a good many missile systems, the coordinate system may be fixed relative to the earth, as the motion of the earth has negligible effects). We therefore define the derivative of a vector V (more properly the inertial derivative) as follows:

If $\vec{V} = x\vec{i} + y\vec{j} + z\vec{k}$, where $\vec{i}$, $\vec{j}$, $\vec{k}$, have fixed directions in inertial space, then

$$(3.1)\quad \dot{\vec{V}} = \dot{x}\vec{i} + \dot{y}\vec{j} + \dot{z}\vec{k} .$$

An alternative definition, equivalent to the above, is

$$(3.2)\quad \dot{\vec{V}}(t) = \lim_{\Delta t \to 0} \frac{\vec{V}(t + \Delta t) - \vec{V}(t)}{\Delta t} .$$

$\dot{\vec{V}}$, being a vector, can be expressed in a second coordinate system by the rules of Part II. Care must be taken that the $\vec{i}$, $\vec{j}$, $\vec{k}$ in the above definition are inertially fixed, for if $\vec{V} = x_1 \vec{i}_1 + y_1 \vec{j}_1 + z_1 \vec{k}_1$, and $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ are in motion, it is not true that $\dot{\vec{V}} = \dot{x}_1 \vec{i}_1 + \dot{y}_1 \vec{j}_1 + \dot{z}_1 \vec{k}_1$. (see exercises below)

Exercises: Prove

(q) $\frac{d}{dt}(\alpha\vec{V}) = \dot{\alpha}\vec{V} + \alpha\dot{\vec{V}}$

(r) $\frac{d}{dt}(\vec{V}_1 + \vec{V}_2) = \dot{\vec{V}}_1 + \dot{\vec{V}}_2$

(s) $\frac{d}{dt}(\vec{V}_1 \cdot \vec{V}_2) = \dot{\vec{V}}_1 \cdot \vec{V}_2 + \vec{V}_1 \cdot \dot{\vec{V}}_2$

(t) $\frac{d}{dt}(\vec{V}_1 \times \vec{V}_2) = \dot{\vec{V}}_1 \times \vec{V}_2 + \vec{V}_1 \times \dot{\vec{V}}_2$

(u) If $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ form a right-handed orthogonal triad of unit vectors, and $\vec{V} = x_1 \vec{i}_1 + y_1 \vec{j}_1 + z_1 \vec{k}_1$, then

$$\dot{\vec{V}} = \dot{x}_1 \vec{i}_1 + \dot{y}_1 \vec{j}_1 + \dot{z}_1 \vec{k}_1 + x_1 \dot{\vec{i}}_1 + y_1 \dot{\vec{j}}_1 + z_1 \dot{k}_1 .$$

Let $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ be a right-handed orthogonal triad of unit vectors, in motion with respect to the inertially fixed triad $\vec{i}$, $\vec{j}$, $\vec{k}$. If one differentiates Eq. (2.2), one obtains

(3.3) $\alpha_{11}\dot{\alpha}_{11} + \alpha_{12}\dot{\alpha}_{12} + \alpha_{13}\dot{\alpha}_{13} = 0$

which is equivalent to

(3.4) $\dot{\vec{i}}_1 \cdot \vec{i}_1 = 0.$

Since then $\dot{\vec{i}}_1$ is normal to $\vec{i}_1$, it can be expressed as a linear combination of $\vec{j}_1$ and $\vec{k}_1$. There then exist functions a and b such that

(3.5) $\dot{\vec{i}}_1 = a\vec{j}_1 + b\vec{k}_1$. Similarly, there exist functions c, d, e, f, such that

(3.6) $\dot{\vec{j}}_1 = c\vec{i}_1 + d\vec{k}_1$.

(3.7) $\dot{\vec{k}}_1 = e\vec{i}_1 + f\vec{j}_1$.

Differentiating Eq. (2.4), the following string of equalities can be seen to hold:

(3.8) $a = \dot{\vec{i}}_1 \cdot \vec{j}_1 = \dot{\alpha}_{11}\,\alpha_{21} + \dot{\alpha}_{12}\,\alpha_{22} + \dot{\alpha}_{13}\,\alpha_{23}$

$= -(\alpha_{11}\,\dot{\alpha}_{21} + \alpha_{12}\,\dot{\alpha}_{22} + \alpha_{13}\,\dot{\alpha}_{23})$

$= -(\dot{\vec{i}}_1 \cdot \dot{\vec{j}}_1) = -c.$

Similarly, it can be shown that $b = -e$ and $d = -f$.

Setting $a = \omega_3$, $e = \omega_2$, $d = \omega_1$, Eqs. (3.5) - (3.7) can be rewritten as

(3.9) $\dot{\vec{i}}_1 = \omega_3\,\vec{j}_1 - \omega_2\,\vec{k}_1$

(3 10) $\dot{\vec{j}}_1 = -\omega_3\,\vec{i}_1 + \omega_1\,\vec{k}_1$

(3.11) $\dot{\vec{k}}_1 = \omega_2\,\vec{i}_1 - \omega_1\,\vec{j}_1$.

If we introduce the vector

(3.12) $\vec{\Omega} = \omega_1\,\vec{i}_1 + \omega_2\,\vec{j}_1 + \omega_3\,\vec{k}_1$

note that where Ω is the angular velocity vector, Eqs. (3.9) - (3.11) are equivalent to

(3.13) $\dot{\vec{i}}_1 = \vec{\Omega} \times \vec{i}_1$.

(3.14) $\dot{\vec{j}}_1 = \vec{\Omega} \times \vec{j}_1$.

(3.15) $\dot{\vec{k}}_1 = \vec{\Omega} \times \vec{k}_1$.

The vector $\vec{\Omega}$ is called the angular velocity vector of the moving coordinate system. Its physical interpretation is that its direction gives the direction about which the coordinate system is instantaneously rotating, its magnitude gives the rate of rotation, the sign of rotation determined by the right-hand rule.

The relationship between the (incorrect) derivative as computed by an individual fixed in a moving coordinate system and the inertial derivative can now be expressed as

$$(3.16)\quad \dot{\vec{V}}_{\text{inertial}} = \dot{\vec{V}}_{\text{moving observer}} + \vec{\Omega} \times \vec{V}.$$

for

$$(3.17)\quad \dot{\vec{V}}_{\text{inertial}} = \dot{x}_1 \vec{i}_1 + \dot{y}_1 \vec{j}_1 + \dot{z}_1 \vec{k}_1 + x_1 (\vec{\Omega} \times \vec{i}_1) + y_1 (\vec{\Omega} \times \vec{j}_1) + z_1 (\vec{\Omega} \times \vec{k}_1)$$

$$= \dot{\vec{V}}_{\text{moving observer}} + \vec{\Omega} \times (x_1 \vec{i}_1 + y_1 \vec{j}_1 + z_1 \vec{k}_1)$$

$$= \dot{\vec{V}}_{\text{moving observer}} + \vec{\Omega} \times \vec{V}.$$

Exercise

(v) Prove

$$\ddot{\vec{V}}_{\text{inertial}} = \ddot{\vec{V}}_{\text{moving observer}} + 2\vec{\Omega} \times \dot{\vec{V}}_{\text{moving observer}} + \dot{\vec{\Omega}} \times \vec{V} + \vec{\Omega} \times (\vec{\Omega} \times \vec{V}).$$

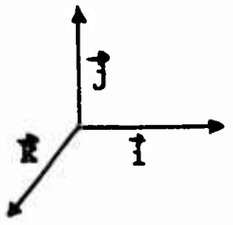
where

$$\ddot{\vec{V}}_{\text{moving observer}} = \ddot{x}_1 \vec{i}_1 + \ddot{y}_1 \vec{j}_1 + \ddot{z}_1 \vec{k}_1.$$

IV. MISSILE MOTION

Let us consider the problem of computing the position of a missile of the cruciform type in space as a function of time. Such a problem would be of importance in a trajectory simulation.

Let $\vec{i}$, $\vec{j}$, $\vec{k}$ be an inertial orthogonal triad, $\vec{j}$ pointing up. (We assume that the earth is flat, and we neglect earth motion).



Let $\vec{P}$ denote the vector joining some fixed point (say the launch site) to the center of gravity of the missile, and let

(4.1) $\vec{P} = x\vec{i} + y\vec{j} + z\vec{k}$.

$\vec{P}$ is called the missile position vector. The motivation of this section is to find P as a function of time.

The missile velocity vector with respect to the launch site is given by

(4.2) $\dot{\vec{P}} = \dot{x}\vec{i} + \dot{y}\vec{j} + \dot{z}\vec{k}$, and the missile acceleration vector by

(4.3) $\ddot{\vec{P}} = \ddot{x}\vec{i} + \ddot{y}\vec{j} + \ddot{z}\vec{k}$.

If $\vec{F}$ is the force on the missile, m its mass, Newton's second law of motion tells

(4.4) $\vec{F} = m\ddot{\vec{P}}$.

Assuming we know $\vec{F}$, we see that it is a simple matter to find $\ddot{x}$, $\ddot{y}$, $\ddot{z}$, and integrating twice gives us our answer. However, $\vec{F}$ comes to us in missile-fixed coordinates. (For example, thrust is along the longitudinal axis, lift normal to it). So a coordinate conversion will be called for.

Let $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ be determined by missile orientation, $\vec{i}_1$ parallel to the longitudinal axis of the missile, $\vec{j}_1$ parallel to one control surface pair, $\vec{k}_1$ parallel to the other.

In general, we will know F_1, F_2, F_3 (more about this later) where Eq.(4.5) $\vec{F} = F_1\vec{i}_1 + F_2\vec{j}_1 + F_3\vec{k}_1 - gm\vec{j}$.

F_1 will be thrust minus drag, F_2 and F_3 aerodynamic lifts. The last term is just gravity.

From Eq. (2.1) we have

$$(4.6)\quad \ddot{x} = \frac{1}{m}\left[F_1\alpha_{11} + F_2\alpha_{21} + F_3\alpha_{31}\right]$$

$$\ddot{y} = \frac{1}{m}\left[F_1\alpha_{12} + F_2\alpha_{22} + F_3\alpha_{32}\right] - g$$

$$\ddot{z} = \frac{1}{m}\left[F_1\alpha_{13} + F_2\alpha_{23} + F_3\alpha_{33}\right] \quad .$$

Integrating twice gives x, y, z. It is clear that α's must be known. That is, we must know the orientation of the missile in space.

If we knew the angular velocity vector $\vec{\Omega}$ of the $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ coordinate system, the α's could be easily found.

For example,

$$(4.7)\quad \dot{\alpha}_{11} = \vec{i}_1 \cdot \dot{\vec{i}} = (\vec{\Omega} \times \vec{i}_1) \cdot \vec{i} = \vec{\Omega} \cdot (\vec{i}_1 \times \vec{i}) = \vec{\Omega} \cdot$$

$$\begin{vmatrix} \vec{i}_1 & \vec{j}_1 & \vec{k}_1 \\ 1 & 0 & 0 \\ \alpha_{11} & \alpha_{21} & \alpha_{31} \end{vmatrix}$$

$$= (\omega_1\vec{i}_1 + \omega_2\vec{j}_1 + \omega_3\vec{k}_1) \cdot (-\alpha_{31}\vec{j}_1 + \alpha_{21}\vec{k}_1)$$

$$= -\omega_2\alpha_{31} + \omega_3\alpha_{21}.$$

Proceeding similarly, we have the equations

$$(4.8)\quad \dot{\alpha}_{11} = -\omega_2 \alpha_{31} + \omega_3 \alpha_{21}$$

$$\dot{\alpha}_{12} = -\omega_2 \alpha_{32} + \omega_3 \alpha_{22}$$

$$\dot{\alpha}_{13} = -\omega_2 \alpha_{33} + \omega_3 \alpha_{23}$$

$$\dot{\alpha}_{21} = \omega_1 \alpha_{31} - \omega_3 \alpha_{11}$$

$$\dot{\alpha}_{22} = \omega_1 \alpha_{32} - \omega_3 \alpha_{12}$$

$$\dot{\alpha}_{23} = \omega_1 \alpha_{33} - \omega_3 \alpha_{13}$$

$$\dot{\alpha}_{31} = -\omega_1 \alpha_{21} + \omega_2 \alpha_{11}$$

$$\dot{\alpha}_{32} = -\omega_1 \alpha_{22} + \omega_2 \alpha_{12}$$

$$\dot{\alpha}_{33} = -\omega_1 \alpha_{23} + \omega_2 \alpha_{13}$$

Integrating these equations then would yield the α's. The numbers ω_1, ω_2, ω_3 can be found as follows if we know the applied moments M_1, M_2, M_3 about the center of gravity which would cause rotations about $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ respectively:

The angular momentum vector is given by

$$(4.9)\quad \vec{H} = I_1 \omega_1 \vec{i}_1 + I_2 \omega_2 \vec{j}_1 + I_2 \omega_3 \vec{k}_1$$

where I_1, I_2, I_2 are the moments of inertia of the missile about $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ respectively, taking the origin at the center of gravity. (For a general rigid body, the expression for $\vec{H}$ is more complicated, involving the products of inertia. In this case, however, because of symmetry, the products of inertia all vanish, and the moments of inertia about $\vec{j}_1$ and $\vec{k}_1$ are equal.)

If we set

(4.10) $\vec{M} = M_1 \vec{i}_1 + M_2 \vec{j}_1 + M_3 \vec{k}_1$, the moment (or torque) vector, Mechanics tells us that

(4.11) $\dot{\vec{H}} = \vec{M}$.

Differentiating Eq.(4.9), we have

$$\begin{aligned}(4.12)\ \dot{\vec{H}} &= I_i \dot{\omega}_1 \vec{i}_1 + I_2 \dot{\omega}_2 j_1 + I_2 \dot{\omega}_3 k_1 + \vec{\Omega} \times H \\ &= I_1 \dot{\omega}_1 \vec{i}_1 + \left[I_2 \dot{\omega}_2 + (I_1 - I_2)\, \omega_1 \omega_3\right] \vec{j}_1 \\ &\quad + \left[I_2 \dot{\omega}_3 + \omega_1 \omega_2 (I_2 - I_1)\right] \vec{k}_1 \\ &= M_1 \vec{i}_1 + M_2 \vec{j}_1 + M_3 \vec{k}_1 = \vec{M}.\end{aligned}$$

Hence

$$(4.13)\ \dot{\omega}_1 = \frac{M_1}{I_1} \text{ (why?)}$$

$$\dot{\omega}_2 = \frac{M_2 - (I_1 - I_2)\, \omega_1\, \omega_3}{I_2}$$

$$\dot{\omega}_3 = \frac{M_3 - (I_2 - I_1)\, \omega_1\, \omega_2}{I_2}$$

Equation (4.13) is a special case of "Euler's equations".

Integrating Eq. (4.13) produces the ω's, which produce the α's, which in turn produce x, y, z. Our problem is now completely solved provided we know F_1, F_2, F_3, M_1, M_2, M_3, m, I_1, I_2. The last three are trivial. The first six are usually obtained from wind tunnel tests. They are generally quite complicated functions of other variables. It is not our purpose to exhibit these functions explicitly, but it will be worthwhile to at least have a look at what these other variables are. F_1 of course depends on thrust. None of the others do, unless there is thrust misalignment. All six depend on

(a) the magnitude and direction of the air stream velocity with respect to the missile,

(b) speed of sound (which changes with temperature),

(c) air pressure (depends on altitude and temperature),

(d) center of gravity position (which moves as fuel is spent) and

(e) the angles δ_1, δ_2, δ_3, δ_4 at which the control surfaces are inclined.

In addition M_1, M_2, M_3 depend on ω_1, ω_2, ω_3 (aerodynamic damping).

A few words about (a) and (e) are in order. If $\vec{W}$ is the velocity vector of the wind with respect to the $\vec{i}$, $\vec{j}$, $\vec{k}$ coordinate system, the $\dot{\vec{P}} - \vec{W}$ is the (negative of the) velocity of the airstream with respect to the missile. Expressing $\dot{\vec{P}} - \vec{W}$ in the $\vec{i}_1$, $\vec{j}_1$, $\vec{k}_1$ coordinate system, we have

(4.14) $\dot{\vec{P}} - \vec{W} = U\,\vec{i}_1 + V\,\vec{j}_1 + W\,\vec{k}_1$, where U, V, W are some numbers.

It is possible to find numbers α, ϕ such that (why?)

(4.15) $\dot{\vec{P}} - \vec{W} = |\dot{\vec{P}} - \vec{W}|\,(\cos\alpha\,\vec{i}_1 + \sin\alpha\sin\phi\,\vec{j}_1 + \sin\alpha\cos\phi\,\vec{k}_1)$.

$|\dot{\vec{P}} - \vec{W}|$ is called air speed, α is called angle of attack, and ϕ is called the aerodynamic roll angle. Air speed divided by speed of sound is called mach number. Data from wind tunnel tests often comes in terms of mach number, angle of attack, and aerodynamic roll angle (other parameters different from these must also be given, such as pressure, for example).

Exercise:

(w) If $\vec{W} = W_1\,\vec{i} + W_2\,\vec{j} + W_3\,\vec{k}$, express air speed, angle of attack, aerodynamic roll angle as functions of W_1, W_2, W_3, $\dot{x}$, $\dot{y}$, $\dot{z}$, α_{11}, α_{12}, α_{13}, α_{21}, α_{22}, α_{23}, α_{31}, α_{32}, α_{33}.

The object of a missile flight is generally to intercept a target. Artistry in many fields is required to achieve this objective in the "best" (a highly subjective term, so we shall use it freely without attempting to define it) possible manner.

The aerodynamic designer tries to make the functional dependencies of F_1, F_2, F_3, M_1, M_2, M_3 on the aforementioned

parameters the best possible.

Given the missile design, we have seen how the missile motion is determined once the missile's environment is known, and once δ_1, δ_2, δ_3, δ_4 are known. δ_1, δ_2, δ_3, δ_4 represent our control over the missile. (Thrust can be controlled in some cases, also). Thus, given the aerodynamic design and the thrust, the problem reduces to: How best should δ_1, δ_2, δ_3, δ_4 be varied in the presence of a varying environment, and how does one do it? This question opens a Pandora's box of problems which today are keeping thousands of people busy. It is not our purpose to continue much further along these lines, although we will subsequently discuss certain aspects which will have some bearing on the matter.

V. TWO DIMENSIONS

We have seen how the equations for a portion of a three-dimensional simulation can be set up. The loop is nowhere near complete, however. The sensing instruments, autopilot, guidance computer, guidance intelligence, as well as other features must also be simulated. These features are discussed elsewhere in the training program.

Such a simulation (the 1103A digital computer at APL is currently engaged in such tasks) is most often used for performance analyses. An autopilot or guidance computer designer will, however, do most of his work in two dimensions, as the three dimensional equations are too complicated to gain insight. Occasionally he will work in three dimensions, when his problems are basically three dimensional in nature. Most of the time this is not necessary.

Accordingly, let us reduce the equations that have been derived so far to two dimensions, where they are of a simpler, more suggestive, and perhaps a more familiar, form.

We select the $\vec{i}$, $\vec{j}$ plane, and assume everything happens in that plane. For simplicity, assume W = 0. Then

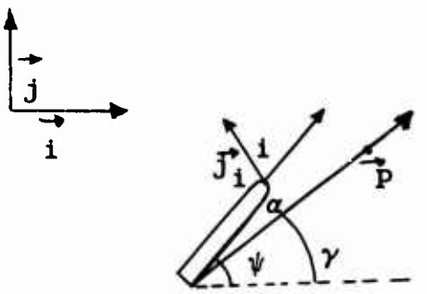
$$(5.1)\quad z = 0,\ F_3 = 0,\ \omega_1 = \omega_2 = 0$$

$$\alpha_{11} = \cos\psi,\ \alpha_{12} = \sin\psi,\ \alpha_{13} = 0$$

$$\alpha_{21} = -\sin\psi,\ \alpha_{22} = \cos\psi,\ \alpha_{23} = 0$$

$\alpha_{31} = 0$, $\alpha_{32} = 0$, $\alpha_{33} = 1$.

$M_1 = M_2 = 0$, $\phi = -\frac{\pi}{2}$, $\delta_1 = \delta_3 = 0$, $\delta_2 = \pm\, \delta_4 = \delta$ (where + or - depends on sign convention)



From the above and Eq.(4.7),

(5.2) $\omega_3 = \dot{\psi}$.

Equation (4.6) becomes

(5.3) $\ddot{x} = \frac{1}{m}[F_1 \cos\psi - F_2 \sin\psi]$

$\ddot{y} = \frac{1}{m}[F_1 \sin\psi + F_2 \cos\psi] - g$.

Equations (4.8) all reduce either to 0 = 0, or Eq. (5.2).

Equations (4.13) reduce to

(5.4) $\ddot{\psi} = \frac{M_3}{I_2}$.

Given the two forces and the moment, Eqs. (5.3) and (5.4) enable us to solve for x and y. (Also, of course, for ψ).

Sometimes it is desirable to write an equivalent set of equations in terms of γ, where if v is missile speed,

(5.5) $v = \sqrt{\dot{x}^2 + \dot{y}^2}$

(5.6) $\dot{x} = v \cos\gamma$

(5.7) $\dot{y} = v \sin \gamma.$

From (4.15),

(5.8) $\dot{\vec{P}} - \dot{\vec{W}} = v \cos \gamma \vec{i} + v \sin \gamma \vec{j} = v \cos \alpha \vec{i}_1 - v \sin \alpha \vec{j}_1$

$= v \cos \alpha (\cos \psi \vec{i} + \sin \psi \vec{j})$

$- v \sin \alpha (- \sin \psi \vec{i} + \cos \psi \vec{j})$

$= v \cos (\psi - \alpha) \vec{i} + v \sin (\psi - \alpha) \vec{j}.$

Hence

(5.9) $\psi = \alpha + \gamma.$

(5.3) becomes

(5.10) $\dot{v} \cos \gamma - v\dot{\gamma} \sin \gamma = \frac{1}{m} [F_1 \cos \psi - F_2 \sin \psi]$

$\dot{v} \sin \gamma + v\dot{\gamma} \cos \gamma = \frac{1}{m} [F_1 \sin \psi + F_2 \cos \psi] - g .$

Multiplying the first equation by cos γ, the second by sin γ, adding, one has

(5.11) $\dot{v} = \frac{1}{m} [F_1'] - g \sin \gamma$

where

(5.12) $F_1' = F_1 \cos \alpha - F_2 \sin \alpha.$

Multiplying the first by - sin γ, the second by cos γ, and adding, one has

(5.13) $v\dot{\gamma} = \frac{1}{m} [F_2'] - g \cos \gamma,$

where

(5.14) $F_2' = F_1 \sin \alpha + F_2 \cos \alpha.$

It is common for an autopilot designer to make further simplifications. He will commonly neglect Eq. (5.11) entirely, assuming that $\dot{v} = 0$. He will assume F_1 and α are small, and replace Eq. (5.14) with $F_2' = F_2 \cos \alpha$. He will neglect gravity, make first order assumptions on $F_2 \cos \alpha$, divide

through by v, and come up with a replacement for Eq. (5.13), namely

(5.15) $\dot{\gamma} = A\alpha + B\delta$

where A and B are functions of mach number, speed of sound, air pressure, center of gravity, mass.

He will linearize Eq. (5.4), coming up with

(5.16) $\ddot{\psi} = -D\dot{\psi} - C\alpha + E\delta$, where C, D, E are functions of the same things A and B are.

Equations (5.9), (5.15), and (5.16) are then used to compute ψ, α, γ as functions of δ. If, in addition, he wishes to know x and y, he uses Eqs. (5.6) and (5.7).

Exercise

(x) Rewrite section V under the additional assumption that $\vec{W}$ lies in the $\vec{i}$, $\vec{j}$ plane, but is different from zero.

VI. MEASUREMENTS MADE BY MECHANICAL END INSTRUMENTS

In the Bumblebee family, there are three basic devices commonly used whose measurements can be simply expressed in vectorial form (when the instruments are perfect, which of course they never are). These are the free gyro, the rate gyro, and the accelerometer. There are other devices used in these and other missiles, such as stable platforms and stapfus. It is beyond the scope of this paper to delve very far into this field. Instead, we shall merely state what the above three devices measure, and discuss some of their applications. We shall not describe the mechanical details of how they do this, leaving this subject for the reader to pursue.

The accelerometer as used in the Bumblebee missiles is mounted in the missile, and is sensitive to accelerations along a chosen direction fixed in the missile. Specifically, if $\vec{n}$ is a unit vector in this direction, the accelerometer measures $(\ddot{\vec{P}} + g\vec{j}) \cdot \vec{n}$, where as before, j points up, and g is acceleration due to gravity. Usually, there are two accelerometers, one for which $\vec{n} = \vec{j}_1$, the other for which $\vec{n} = \vec{k}_1$. They are used in the missile autopilot. The acceleration command is compared to the accelerometer output, and the wings are driven until the difference is zero. An autopilot which functions in

this manner is called an accelerometer (or acceleration) feedback autopilot.

The rate gyro is sometimes, but not invariably, mounted in the missile, and is sensitive to rotations about a chosen direction fixed in the missile. Specifically, if n is a unit vector in this direction, the rate gyro measures $\vec{\Omega} \cdot \vec{n}$.

Usually, there are three rate gyros, for $\vec{n} = \vec{i}_1, \vec{j}_1, \vec{k}_1$, respectively. The latter two are used for damping in the autopilot, and the former in the roll control loop, One will also find rate gyros mounted on homing seekers, whose purpose is to aid in space stabilization of these devices, as well as to furnish steering error signals to the guidance computer.

The free gyro has two degrees of freedom, and two possible outputs. The free gyro is used for attitude stabilization during boost, where both outputs are used, and in the roll control loop during beam riding, where only one output is used. There are three definitive directions, denoted by the unit vectors $\vec{S}$, $\vec{G}_O$, $\vec{G}_I$, where S is the direction of the spin axis, fixed in inertial space by virtue of the action of the gyro. $\vec{G}_I$ is the inner gimbal axis, and is always perpendicular to both $\vec{S}$ and $\vec{G}_O$. $\vec{G}_O$ is the outer gimbal axis, fixed in the missile. Let n be some unit vector fixed in the missile which is perpendicular to $\vec{G}_O$. The two possible outputs can be expressed in terms of $\vec{S}$, $\vec{G}_O$, n since clearly $\vec{G}_I = \dfrac{\vec{S}\times\vec{G}_O}{|\vec{S}\times\vec{G}_O|}$.

(6.1) $\cos\theta_1 = \vec{S} \cdot \vec{G}_O$.

The second output is the angle* θ_2 between G_I and n.

(6.2) $\cos\theta_2 = \vec{n} \cdot \dfrac{\vec{S}\times\vec{G}_O}{|\vec{S}\times\vec{G}_O|}$.

In the roll control loop, $\vec{S}$ is usually horizontal, $\vec{G}_O$ coincident with $\vec{i}_1$. Only the angle θ_2 is used.

In the attitude stabilization loop for boost, $\vec{G}_O$ is $\vec{j}_1$, $\vec{n}$ is $\vec{k}_1$, $\vec{S}$ is in the direction of firing.

Exercise

(y) Prove that $\sin\theta_2 = \pm \dfrac{\vec{n} \cdot \vec{S}}{|\vec{S} \times \vec{G}_O|}$, the sign depending on convention.

*or some function of it.

VII. OTHER USEFUL COORDINATE SYSTEMS

In this paper, most of the useful techniques for manipulating coordinate systems have been presented. Practically all three dimensional work with missile motion can be handled by a judicious use of the concepts and basic formulae developed here. (There is, of course, nothing here that is new, and everything can be found in one or more well known texts on Mechanics or Vector Analysis)

Having developed the basic concepts and formulae, a few applications and practical manipulations have been presented. We could go on at considerable length discussing other applications and special coordinate systems in detail. This will not be done here, however, as any reader will now have little difficulty in mastering such topics, should he come in direct contact with them.

We will, however, mention in passing some other useful coordinate systems.

Wind-Fixed Coordinate Systems---It sometimes simplifies calculations when a constant velocity wind is being studied to fasten the tail of the vector $\vec{P}$ to a point moving with the wind. This is still a Newtonian frame of reference, so that Eqs. (4.4) and (4.11) hold. In this coordinate system, Section V carries over intact. (Compare with the results of Exercise (x).)

Radar-Fixed Coordinate Systems---Here the triad of unit vectors is fixed in the dish of a tracking radar or a guided missile guidance transmitter. This is useful when studying tracking radar dynamics or certain beam riding problems, in particular for developing expressions for beam riding error signals.

Deck-Fixed Coordinate Systems---When missiles are being fired from a ship, the motion of the ship presents new problems from the land-based case. Since radars are normally attached to the deck, their measurements are most easily obtained in a deck-fixed coordinate system. Stabilization of these radars is then required. This calls for coordinate conversions from deck to radar, or deck to inertial, of one type or another. These conversions are accomplished physically by the judicious use of gyroscopic devices, such as rate gyros or gyrocompasses. (The latter is a form of stable platform).

Homing Seeker-Fixed Coordinate Systems---A homing seeker is a tracking radar mounted in the missile, and properly choosing certain unit vectors fixed in the seeker aids in studying seeker dynamics.

Inertial Guidance Coordinate Systems---In inertially guided aircraft or (long range) missiles, it is sometimes convenient to choose a moving coordinate system, one of whose unit vectors points in the direction of the normal to the earth's surface at the position of the aircraft or missile. Sometimes people choose instead the direction of gravity, and sometimes the direction to the center of the earth. (These three directions are slightly different).

The initial distribution list of this document has been made in accordance with a list on file in the Technical Reports Group of the Johns Hopkins University, Applied Physics Laboratory.